NCS:066

Data Warehousing and Data Mining

Shikha Gautam
Asst.Professor

Topic Covered

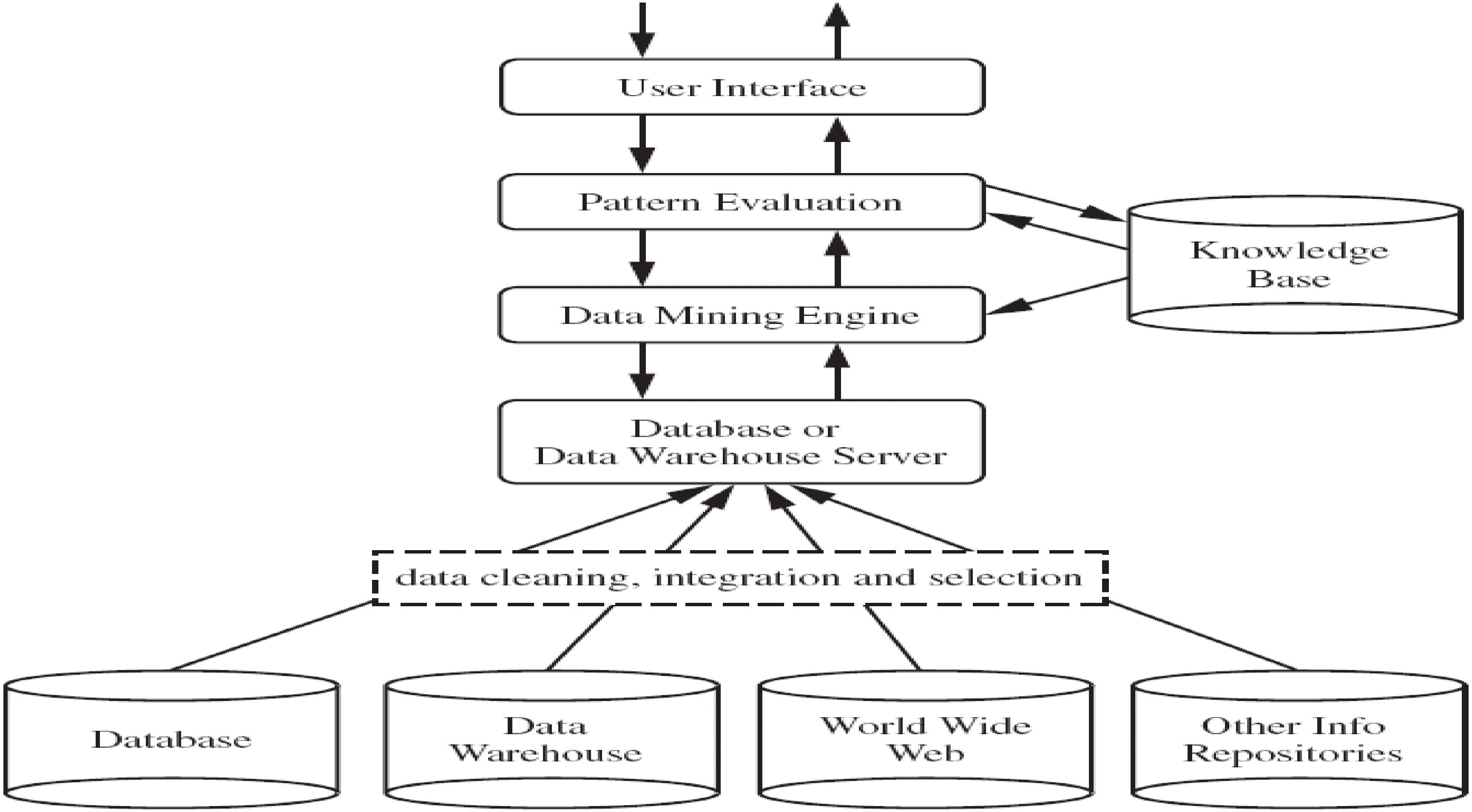
- Introduction ,
- Motivation(Data Mining),
- KDD Process

What Is Data Mining??

- Data mining is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases.

- Alternative names:
  - Knowledge discovery(mining) in databases (KDD),
  - knowledge extraction,
  - data/pattern analysis,
  - data archeology,
  - business intelligence, etc.

A typical DM System Architecture (2)



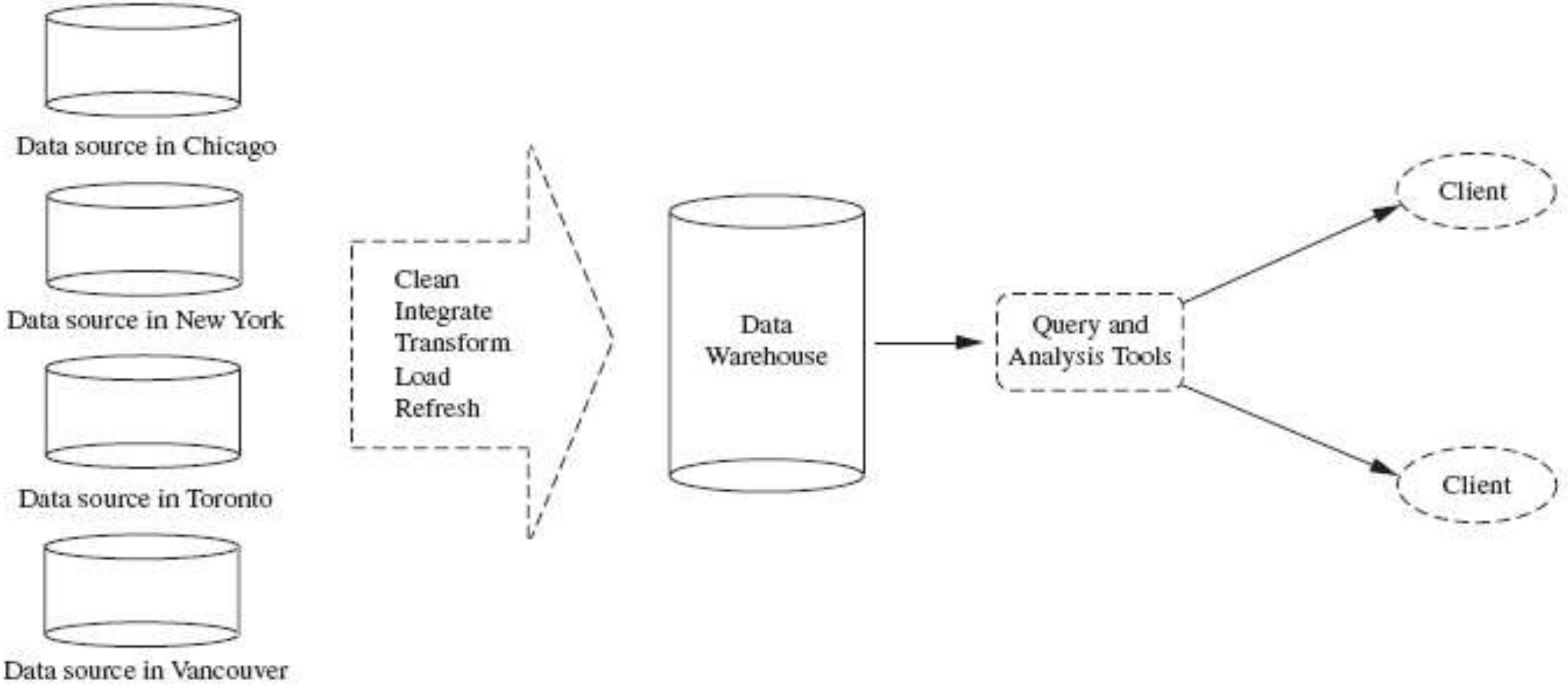
On What Kinds of Data?

- **Database-oriented data sets and applications**
  - Relational database, data warehouse, transactional database

- **Advanced data sets and advanced applications**
  - Object-Relational Databases
  - Temporal Databases
  - Spatial Databases and Spatiotemporal Databases
  - Text databases and Multimedia databases
  - Heterogeneous Databases
  - Data Streams
  - The World-Wide Web etc.

Data Warehouses

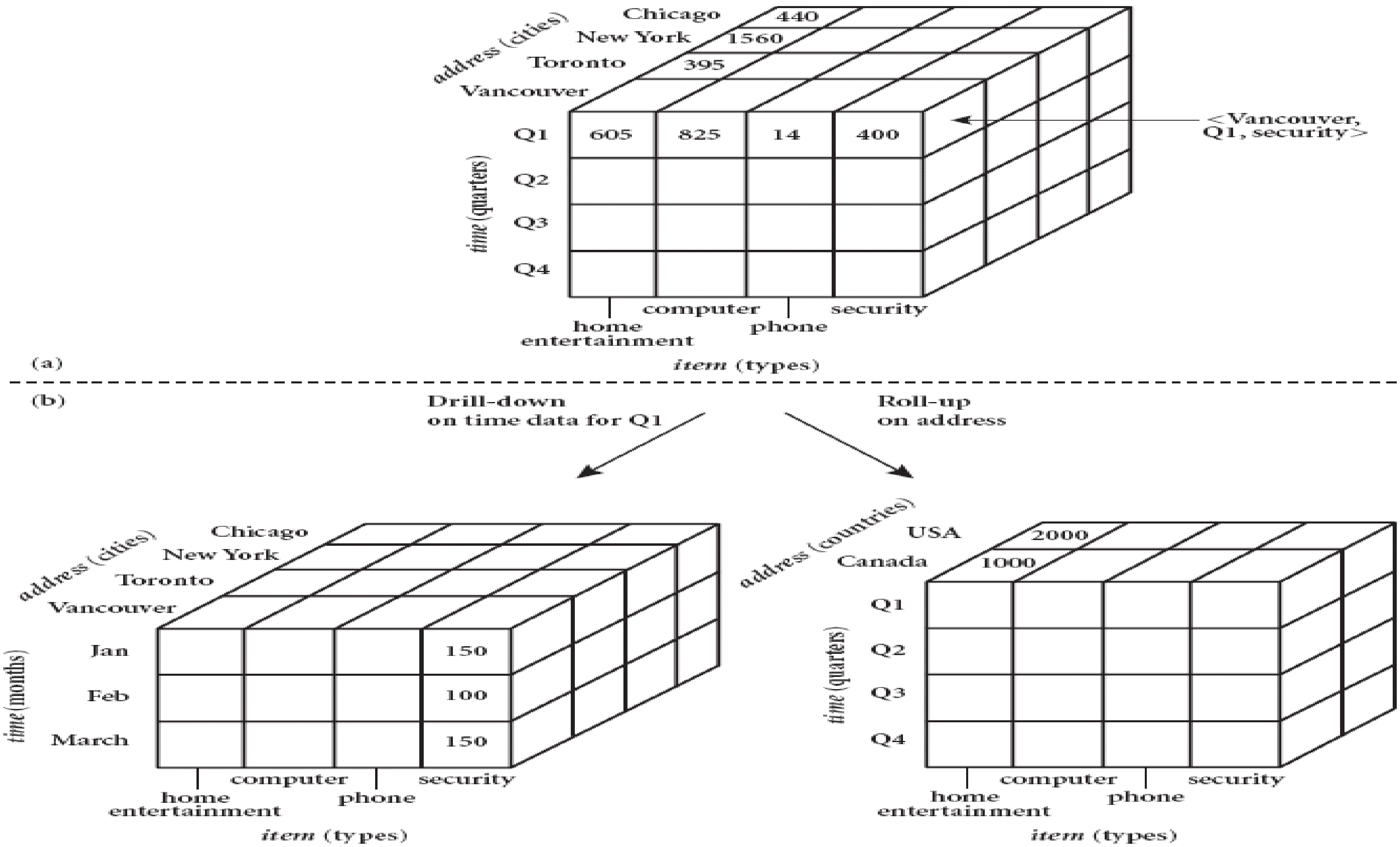
- A repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
- Constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.



Data Warehouses (2)

- Data are organized around major subjects, e.g. customer, item, supplier and activity.
- Provide information from a historical perspective (e.g. from the past 5 – 10 years)
- Typically summarized to a higher level (e.g. a summary of the transactions per item type for each store)
- User can perform drill-down or roll-up operation to view the data at different degrees of summarization

Data Warehouses (3)

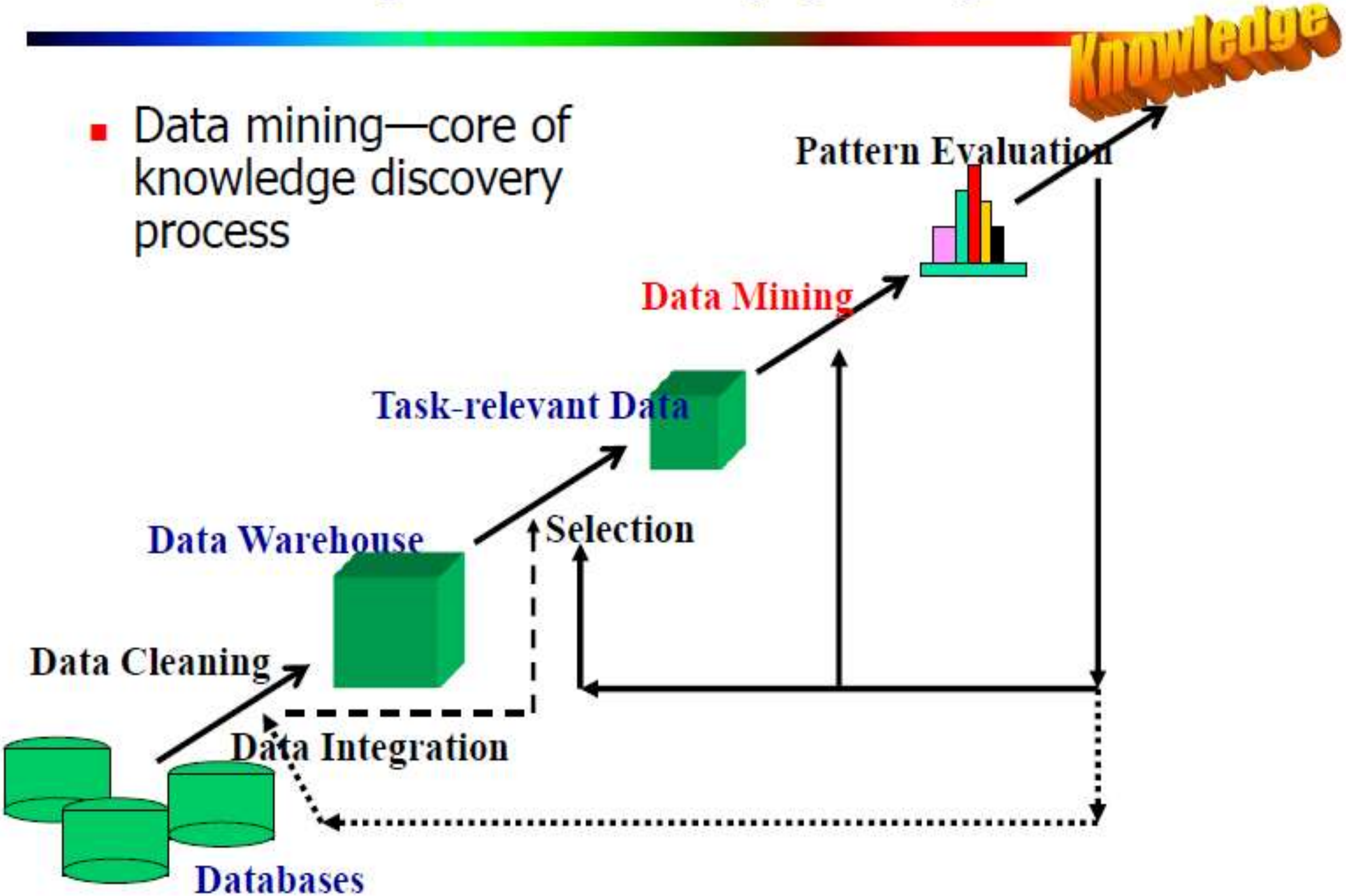


Applications of Data Mining

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis

Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process



KDD Process: Several Key Steps

1. Learning the application domain.

2. Identifying a target data set: data selection

3. Data processing

- **Data cleaning** (remove noise and inconsistent data)
- **Data integration** (multiple data sources maybe combined)
- **Data selection** (data relevant to the analysis task are retrieved from database)
- **Data transformation** (data transformed or consolidated into forms appropriate for mining)

  (Done with data preprocessing)
- **Data mining** (an essential process where intelligent methods are applied to extract

  data patterns)
- **Pattern evaluation** (indentify the truly interesting patterns)
- **Knowledge presentation** (mined knowledge is presented to the user with visualization or representation techniques)

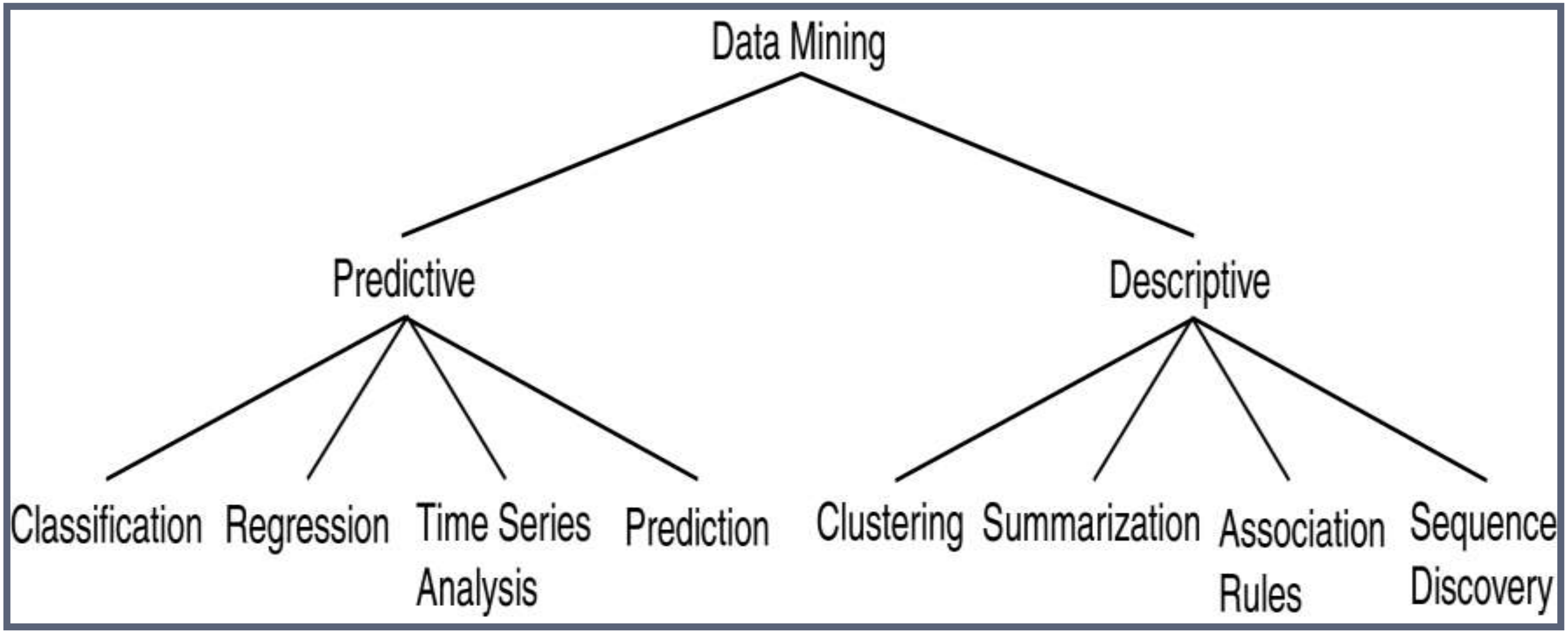
4. **Use of discovered knowledge**

Data mining functionalities-

1. Descriptive Data mining Tasks: Describe general properties of existing data.

2. Predictive Data mining Tasks: Attempt to do predictions based on inference on available data.

Data Mining Functionalities/Tasks



Topic Covered

- Characterization,
- Discrimination ,
- Association,
- Classification,
- Prediction,
- Clustering,
- Outlier analysis

Classification

Classification is a process of predicting class label for unseen new data based on the data tuples with known class labels

Examples:

- ➢Predict whether a new customer buy a Computer in the store ?
- ➢Predict the loan applicant status as safe or risky

Classification(Cont'd)

Classification is a 2 step process

Step-1: **Construction or training of a model/ classifier** using training data tuples with known class labels

Step-2: **Testing the accuracy of a classifier** using the test data tuples for which the class label is already known

Classification Algorithms

- Decision tree-Induction(Tree like flowchat)
- Back propagation (Neural Network)
- Bayesian Classification (statistical probability)
- Rule-based classification (If..Else rules)

Classification is called as **supervised-Learning**

1. Characterization

- Summarization of general features of objects in a target class, and produces what is called Characteristic rules.

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

- Concept/Class Description: Characterization and Discrimination
  - Descriptions can be derived via
    - Data characterization – summarizing the general characteristics of a

      target class of data.
      - E.g. summarizing the characteristics of customers who spend more than $1,000 a year

        at *AllElectronics*. Result can be a general profile of the customers, such as 40 – 50 years old, employed, have excellent credit ratings.

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

    - Data discrimination – comparing the target class with one or a set of comparative classes
        - E.g. Compare the general features of software products whole sales increase by 10% in the last year with those whose sales decrease by 30% during the same period
    - Or both of the above

- Mining Frequent Patterns, Associations and Correlations
    - Frequent itemset: a set of items that frequently appear together in a transactional data set (e.g. milk and bread)
    - Frequent subsequence: a pattern that customers tend to purchase product A, followed by a purchase of product B

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

- Association Analysis: find frequent patterns
  - E.g. a sample analysis result – an association rule:

    buys(X, “computer”) => buys(X, “software”) [support = 1%, confidence = 50%]

    (if a customer buys a computer, there is a 50% chance that she will buy software. 1% of all of the transactions under analysis showed that computer and software are purchased together.)
  - Associations rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold.
- Correlation Analysis: additional analysis to find statistical correlations between associated pairs

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

- Classification and Prediction
  - Classification
    - The process of finding a model that describes and distinguishes the data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
    - The derived model is based on the analysis of a set of training data (data objects whose class label is known).
    - The model can be represented in *classification (IF-THEN) rules*, decision trees, *neural networks*, etc.
  - Prediction
    - Predict missing or unavailable numerical data values

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

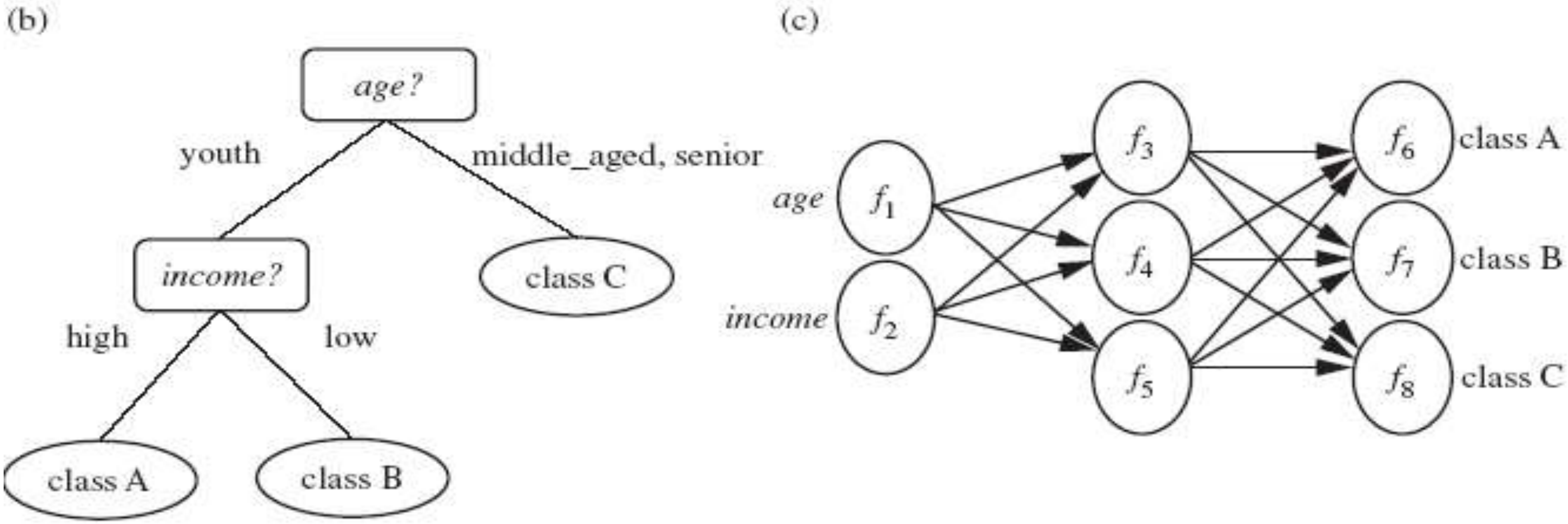
(a)

age(X, "youth") AND income(X, "high") → class(X, "A")

age(X, "youth") AND income(X, "low") → class(X, "B")

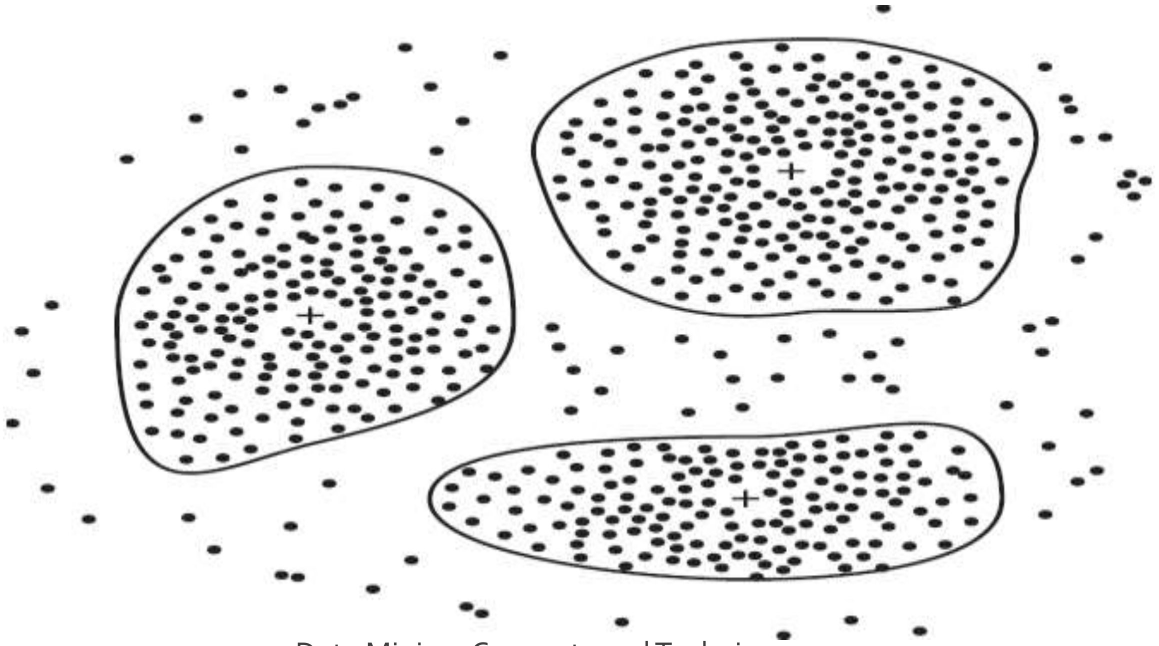
age(X, "middle_aged") → class(X, "C")

age(X, "senior") → class(X, "C")



Data Mining Functionalities (2)

- Cluster Analysis
  - Class label is unknown: group data to form new classes
  - Clusters of objects are formed based on the principle of *maximizing intra-class similarity* & *minimizing interclass similarity*
    - E.g. Identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.



Data Mining Functionalities (2)

- Outlier Analysis
  - Data that do no comply with the general behavior or model.
  - Outliers are usually discarded as noise or exceptions.
  - Useful for fraud detection.
    - E.g. Detect purchases of extremely large amounts
- Evolution Analysis
  - Describes and models regularities or trends for objects whose behavior changes over time.
    - E.g. Identify stock evolution regularities for overall stocks and for the stocks of particular companies.

Data Mining Tasks/Functionalities

- Prediction Tasks
  - Use some variables to predict unknown or future values of other variables
- Description Tasks
  - Find human-interpretable patterns that describe the data.

Common data mining tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

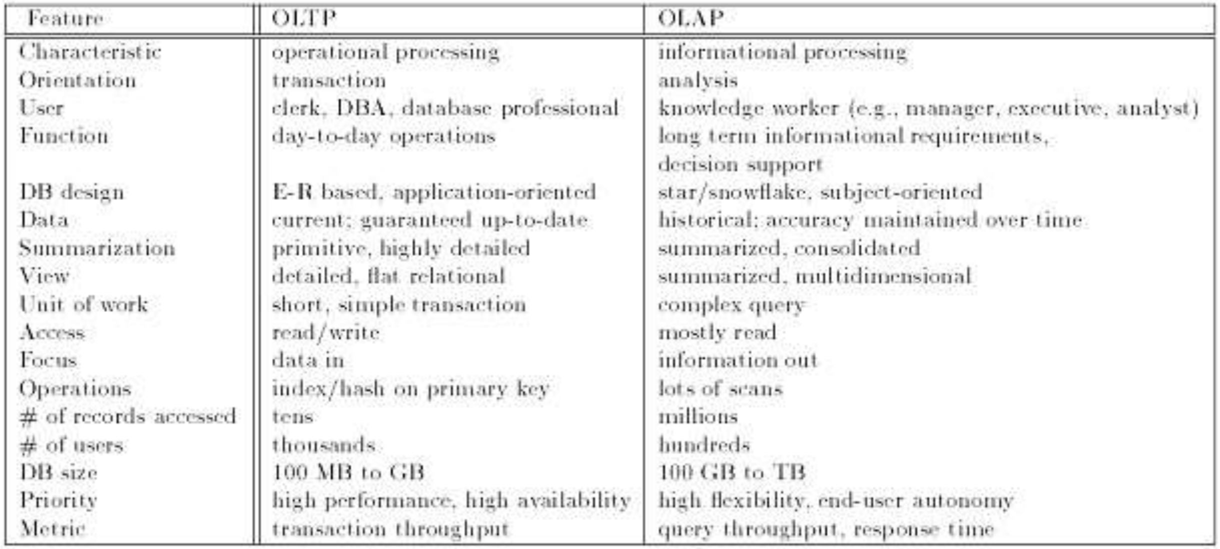
Database Processing vs. Data Mining Processing

- Query
  - Well defined
  - SQL
- Data
  - – Operational data
- Output
  - – Precise
  - – Subset of database

- Query
  - Poorly defined
  - No precise query language
- Data
  - – Not operational data
- Output
  - – Fuzzy
  - – Not a subset of database

Query Examples

- Database
  - – Find all credit applicants with last name of Smith.
  - – Identify customers who have purchased more than $10,000 in the last month.
  - – Find all customers who have purchased milk
- Data Mining
  - – Find all credit applicants who are poor credit risks. (classification)
  - – Identify customers with similar buying habits. (Clustering)
  - – Find all items which are frequently purchased with milk. (association rules)



Feature	OLTP	OLAP
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long term informational requirements, decision support
DB design	E-R based, application-oriented	star/snowflake, subject-oriented
Data	current; guaranteed up-to-date	historical; accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
# of records accessed	tens	millions
# of users	thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Classification

Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

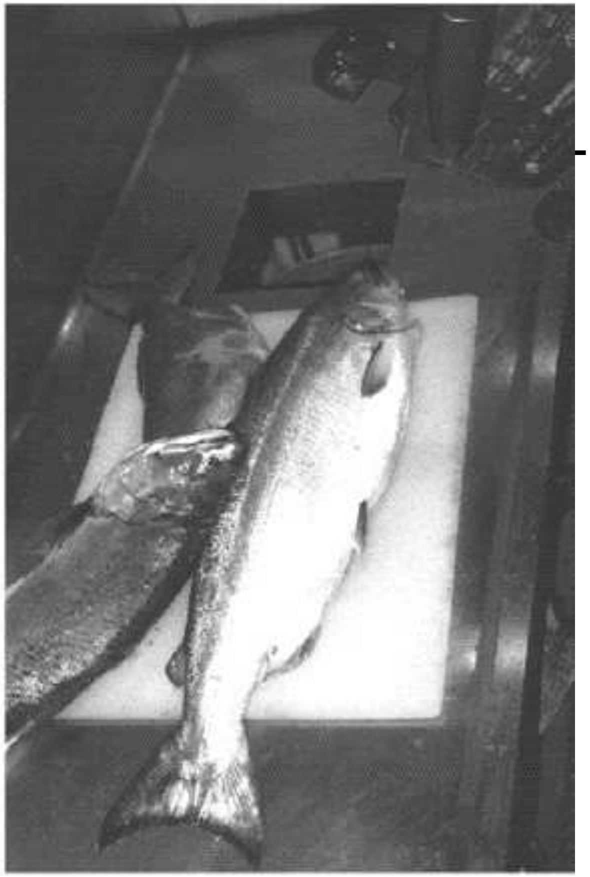
An Example

(from *Pattern Classification by* Duda & Hart & Stork – Second Edition, 2001)

- A fish-packing plant wants to automate the process of sorting incoming fish according to species

- As a pilot project, it is decided to try to separate sea bass from salmon using optical sensing

An Example (continued)



Features (to distinguish):

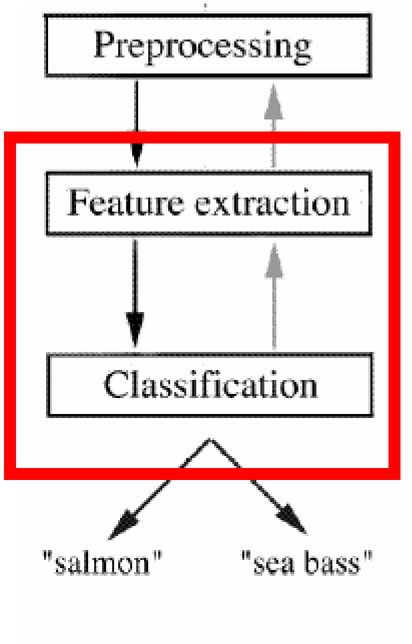
Length

Lightness

Width

Position of mouth

An Example (continued)



- **Preprocessing:** Images of different fishes are isolated from one another and from background;
- **Feature extraction:** The information of a single fish is then sent to a feature extractor, that measure certain “features” or “properties”;
- **Classification:** The values of these features are passed to a classifier that evaluates the evidence presented, and build a model to discriminate between the two species

An Example (continued)

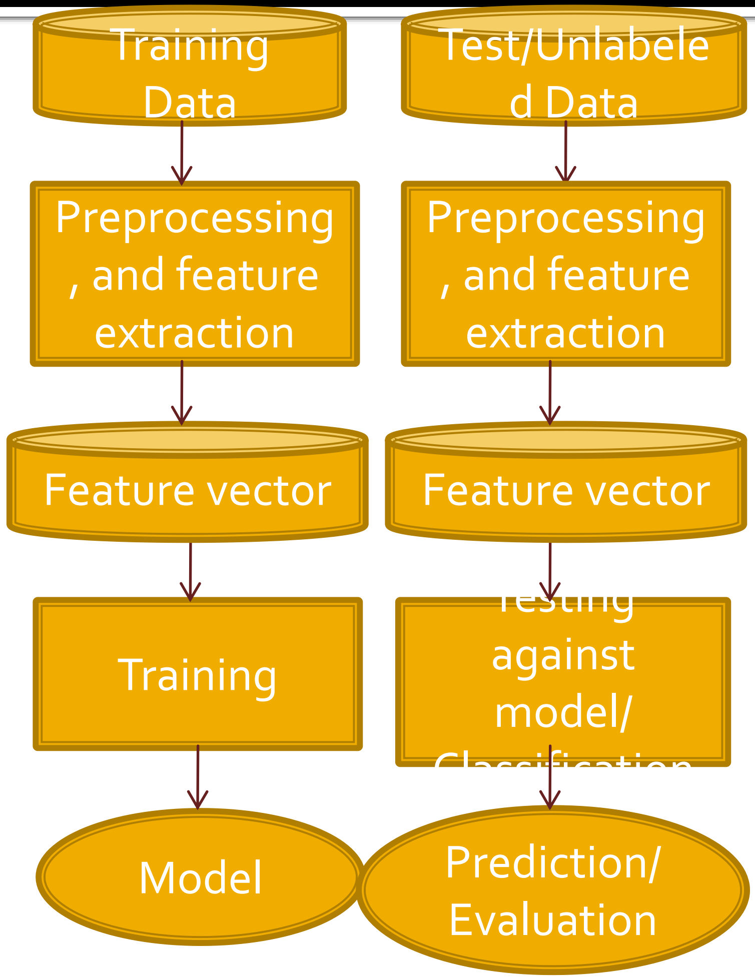
- Domain knowledge:
  - A sea bass is generally longer than a salmon
- Related feature*:* (or attribute)
  - Length
- Training the classifier:
  - Some examples are provided to the classifier in this form: <fish_length, fish_name>
  - These examples are called training examples
  - The classifier *learns* itself from the training examples, how to distinguish Salmon from Bass based on the *fish_length*

An Example (continued)

- Classification model (hypothesis):
  - The classifier generates a model from the training data to classify future examples (test examples)
  - An example of the model is a rule like this:
  - If *Length >= l* then sea bass* otherwise *salmon*
  - Here the value of *l** determined by the classifier
- Testing the model
  - Once we get a model out of the classifier, we may use the classifier to test future examples
  - The test data is provided in the form <fish_length>
  - The classifier outputs <fish_type> by checking *fish_length* against the model

An Example (continued)

- So the overall classification process goes like this →



Training Data

Test/Unlabeled Data

Preprocessing, and feature extraction

Preprocessing, and feature extraction

Feature vector

Feature vector

Training

Testing against model/ Classification

Model

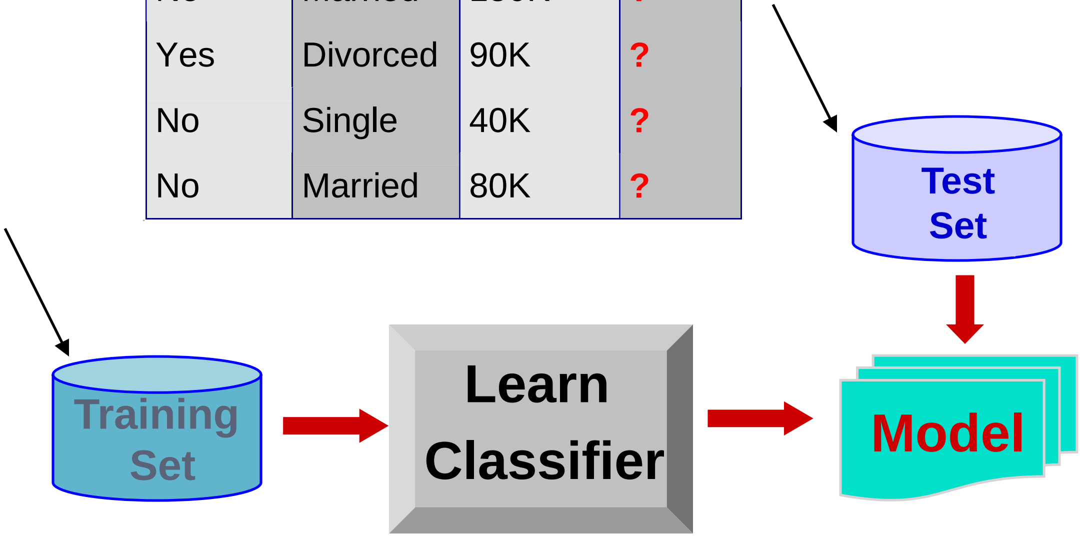
Prediction/ Evaluation

Classification Example 2

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Clustering

Clustering Definition

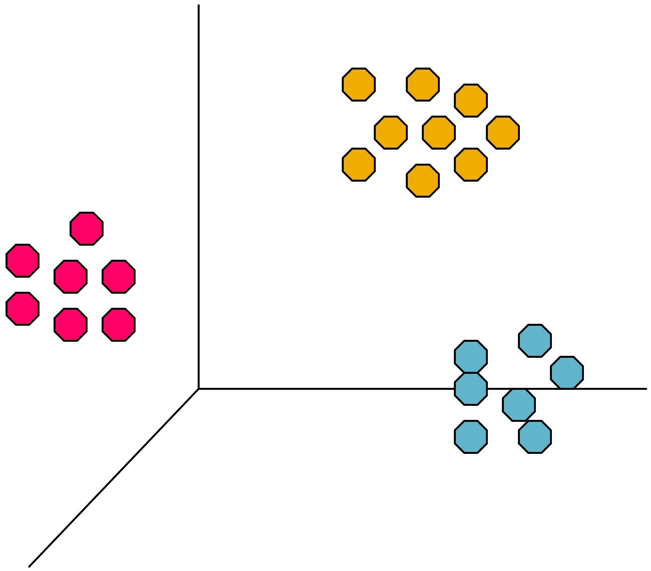
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Association rule mining

Association Rule Discovery: Definition

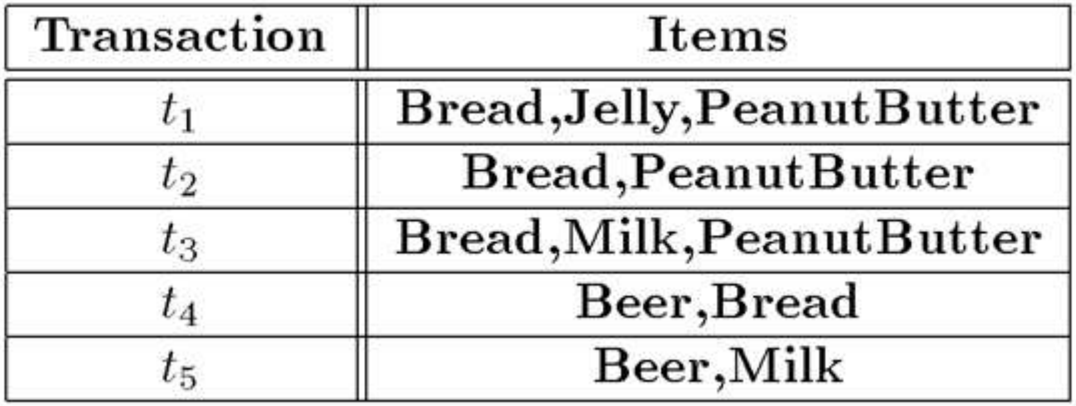
- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Rules Discovered:

{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

- Association Rules
  - Implication: $X \rightarrow Y$ where $X,Y \subseteq I$ and $X \cap Y = \varnothing$;
  - Support of AR (s) $X \rightarrow Y$:
    - Percentage of transactions that contain
- $X \cup Y$
    - Probability that a transaction contains $X \cup Y$.
-
  - Confidence of AR (a) $X \rightarrow Y$:
    - Ratio of number of transactions that contain $X \cup Y$ to the number that contain X
    - Conditional probability that a transaction having X also contains *Y*.



Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

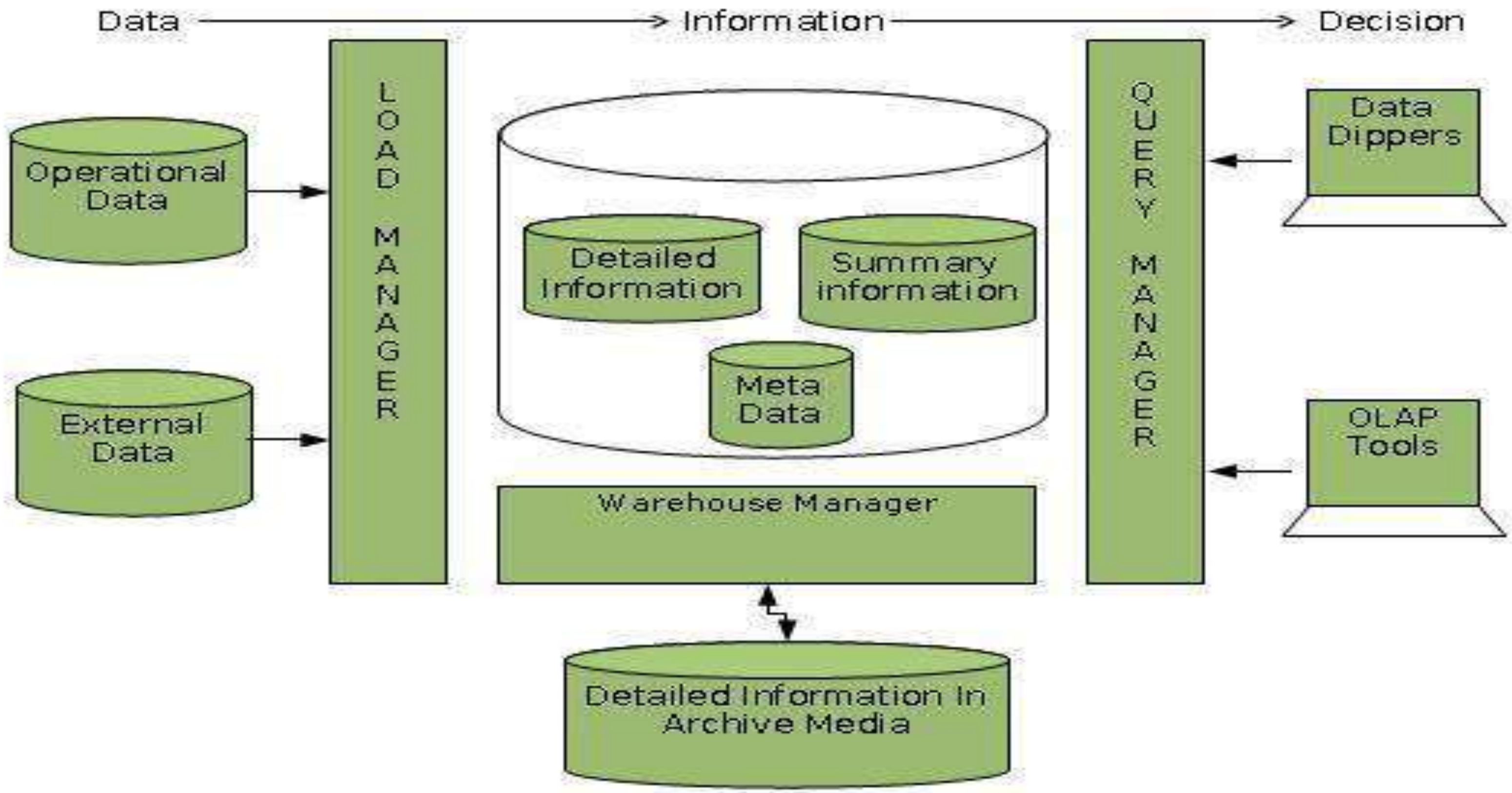
Bre		
Peanutbutter ➔ Bread	60%	= (3/5)/(3/5)%=100%
Jelly ➔ Milk	0%	0%
Jelly ➔ Peanutbutter	=1/5 % = 20%	= (1/5)/(1/5) % = 100%

Usage of data warehouse

- Three kinds of data warehouse applications: information processing, analytical processing, and data mining:
- **Information processing**: supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts or graphs.

- **Analytical processing**: supports basic OLAP operations, including slice-and-dice, drill-down, roll-up, and pivoting.
- **Data mining**: supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

Data warehouse Components



Load Manager

- This component performs the operations required to extract and load process.
- The size and complexity of the load manager varies between specific solutions from one data warehouse to other.

- The load manager performs the following functions:
  - The load manager performs the following functions:
  - Extract the data from source system.
  - Fast Load the extracted data into temporary data store.
  - Perform simple transformations into structure similar to the one in the data warehouse.

Warehouse Manager

- A warehouse manager is responsible for the warehouse management process. It consists of third-party system software, C programs, and shell scripts.
- The size and complexity of warehouse managers varies between specific solutions.

- A warehouse manager includes the following:
  - The controlling process
  - Stored procedures or C with SQL
  - Backup/Recovery tool
  - SQL Scripts

Query Manager

- Query manager is responsible for directing the queries to the suitable tables.
- By directing the queries to appropriate tables, the speed of querying and response generation can be increased.
- Query manager is responsible for scheduling the execution of the queries posed by the user.

Topic Covered

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization

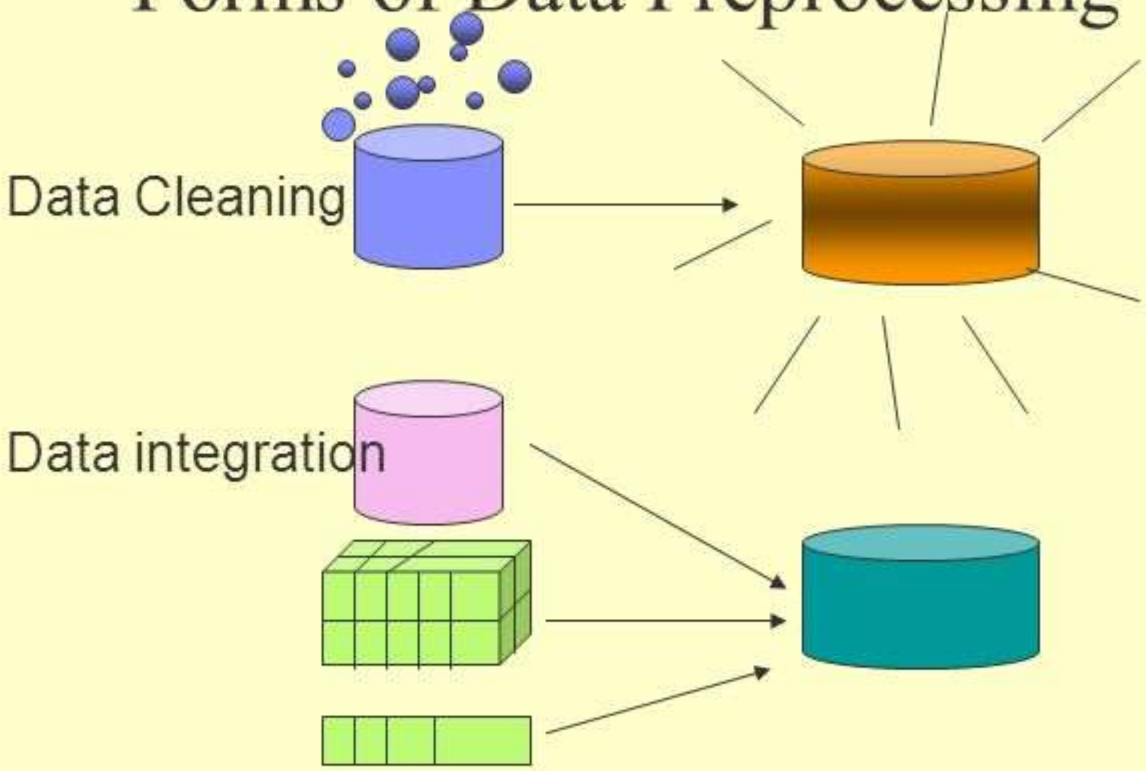
Why Preprocess the Data?

- Data in the real world is dirty.
- No quality data, no quality mining results!
- A multi-dimensional measure of data quality.
  - accuracy, completeness, consistency, timeliness, believability.

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization (scaling to a specific range)
  - Aggregation

Forms of Data Preprocessing

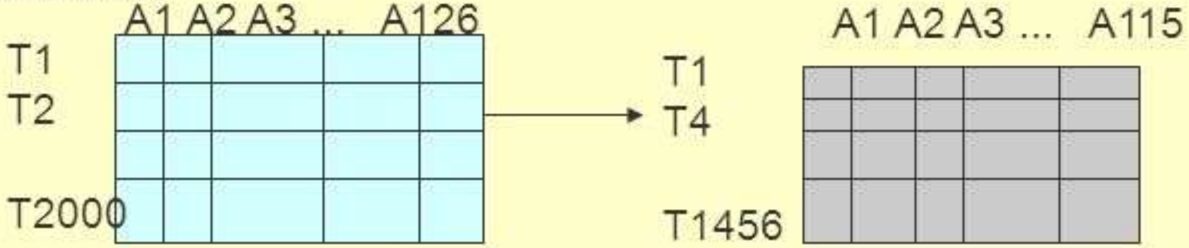


Data Cleaning

Data integration

Data transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data reduction



A1 A2 A3 ... A126

T1
T2
T2000

A1 A2 A3 ... A115

T1
T4
T1456

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. e.g., *Occupation*="" " (missing data)
  - Noisy: containing noise, errors, or outliers. e.g., *Salary*="−10" (an error)
  - Inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*="42", *Birthday*="03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

Tasks of Data Cleaning

- Fill in missing values
- Identify outliers and smooth noisy data
- Correct inconsistent data

Data Cleaning Approaches

- **Data Analysis:** Detect which kinds of errors and inconsistencies are to be removed.

- **Definition of transformation workflow and mapping rules:**

- **Verification:** Test correctness and effectiveness of transformation workflow and Transformation definition.

Data Cleaning Approaches

- **Transformation:** Execution of transformation steps either by running ETL workflow for loading or during answering queries on multiple sources.

- **Backflow of cleaned data:** Cleaned data should replace dirty data.

Manage Missing Data

- **Ignore the tuple:** usually done when class label is missing.
- **Fill in the missing value manually:** tedious + infeasible?
- **Use a global constant to fill in the missing value:** e.g., "unknown", a new class?!
- **Use the attribute mean to fill in the missing value**
  Use the attribute mean for all samples of the same class to fill in the missing value: smarter
- **Use the most probable value to fill in the missing value:** inference based such as regression, Bayesian formula, decision tree

Noisy Data

“Noise is a random error or variance in a measured variable.”

- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

Manage Noisy Data

- Binning Method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc
- Clustering:
  - detect and remove outliers
- Semi Automated
  - Computer and Manual Intervention
- Regression
  - Use regression functions

Manage Inconsistant Data

- Manual correction using external references.
- Semi-automatic using various tools:

  - To detect violation of known functional dependencies and data constraints.
  - To correct redundant data.

NCS:066

KNOW YOUR DATA

Shikha Gautam
Asst.Professor
PSIT coe (CSE)

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Data objects are described by **attributes**.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples* , *examples*, *instances*, *data points*, *objects*, *tuples*.
- Database rows -> data objects; columns ->attributes.

Attributes

- **Attribute** (or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*

Attribute Types

- **Nominal Attribute:** Also known as Categorical.
  - The values of nominal attribute are symbols or name of things.
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval-Scaled Attributes:**
    - Measured on a scale of **equal-sized units**
    - Values have order
        - E.g., *temperature in C°or F°, calendar dates*
    - No true zero-point
- **Ratio-Scaled Attributes:**
    - Inherent **zero-point**
    - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).
        - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Attribute Types (Cont'd)

- **<u>Binary Attribute</u>:**
  - Nominal attribute with only 2 states (0 and 1)
  - <u>Symmetric binary</u>: both outcomes equally important
    - e.g., gender
  - <u>Asymmetric binary</u>: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

- **<u>Ordinal Attribute:</u>**
  - Values have a meaningful order (ranking) .
  - *Size* = {*small*, *medium*, *large*}, grades, army rankings

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents, age
  - Note: Binary attributes are a special case of discrete attributes

- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Continuous attributes are typically represented as floating-point variables

Basic Statistical Descriptions of Data

To better understand the data: central tendency, variation and spread

- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.

Measuring the Central Tendency

- **Mean (algebraic measure):** Numeric Measure of the "center" of a set of data is the (arithmetic) mean.

  - Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Trimmed mean: chopping extreme values

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Measuring the Central Tendency (Cont'd)

- **Median:**
  - Middle value if odd number of values,
  - Average of the middle two values ,Otherwise.
  - Estimated by interpolation (for *grouped data*):

$$median = L_1 + \left(\frac{n/2 - (\sum freq)l}{freq_{median}}\right) width$$

Measuring the Central Tendency (Cont'd)

- **Mode:**
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

$$mean - mode = 3 \times (mean - median)$$

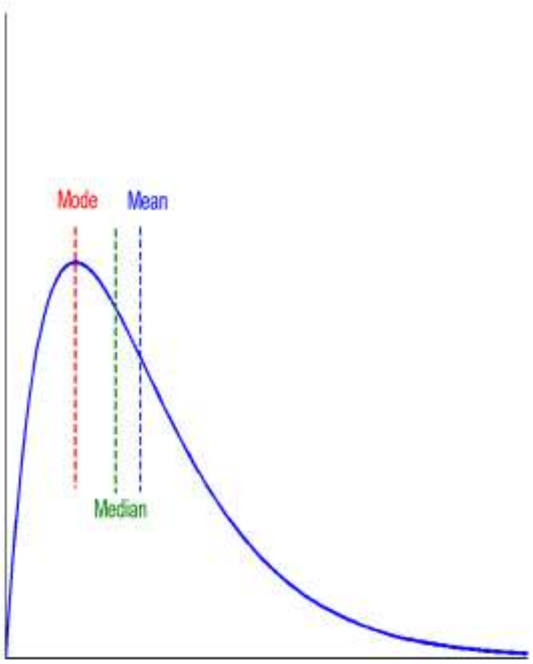
- **Midrange**:

(min_value + max_value)/2

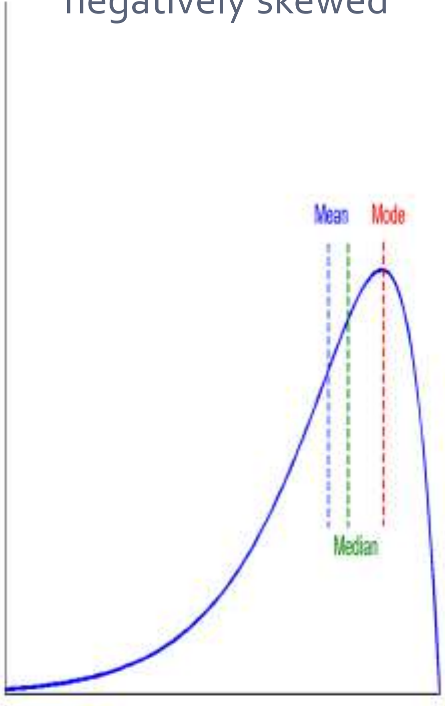
Symmetric vs. Skewed Data

Median, mean and mode of symmetric, positively and negatively skewed data.

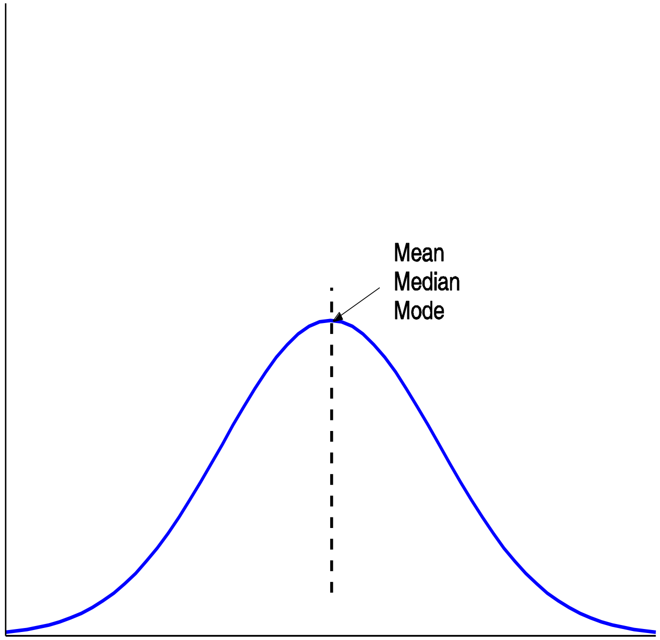
positively skewed



negatively skewed



symmetric



Measuring the Dispersion of Data

- **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
- **Inter-quartile range**: $IQR = Q_3 - Q_1$
- **Five number summary**: min, Q_1, median, Q_3, max
- **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
- **Outlier**: usually, a value higher/lower than 1.5 x IQR

Measuring the Dispersion of Data (Cont'd)

- **Variance**:

How spread out a data distribution is.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i{}^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i{}^2 - \mu^2$$

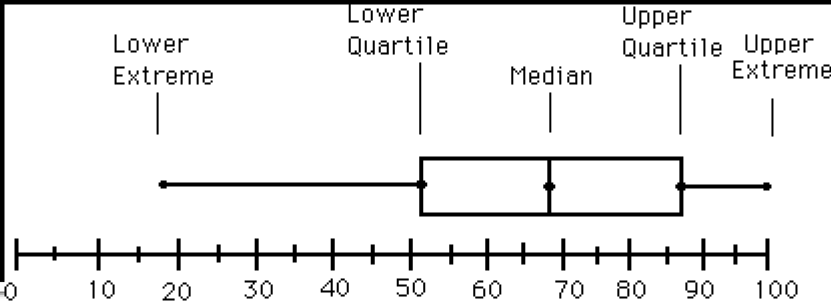
Measuring the Dispersion of Data (Cont'd)

- **Standard deviation**:

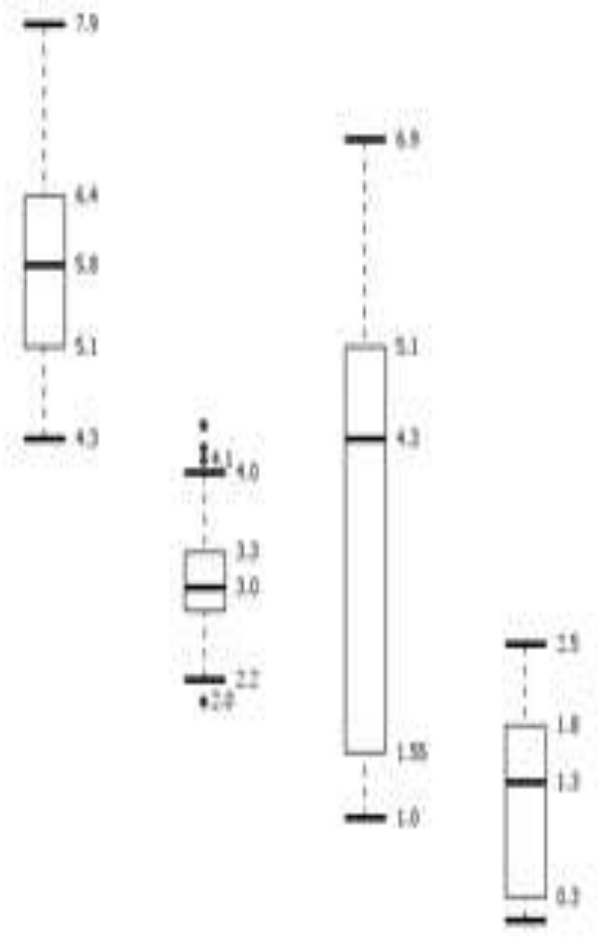
s *(or σ)* is the square root of variance s^2 *(or* σ^2*)*

  - Low standard deviation means that the data observation tend to be very close to the mean.
  - High standard deviation means that the data are spread out over a large range of values.

Boxplot Analysis



- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - **Whiskers**: two lines outside the box extended to Minimum and Maximum
  - **Outliers:** points beyond a specified outlier threshold, plotted individually



Graphic Displays of Basic Statistical Descriptions

- Methods for visual inspection of data:
  - Quantile plot
  - Quantile- Quantile plot
  - Histogram

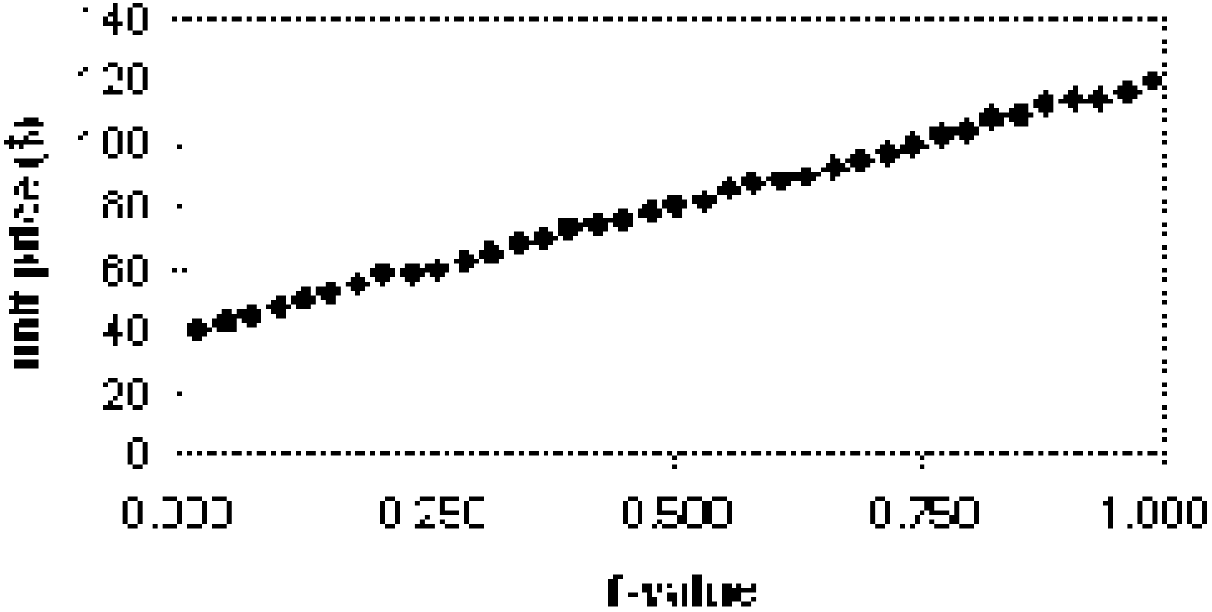
Univariate distribution (Data for one attribute).

  - Scatter Plot

Bivariate distribution (Involving two attributes).

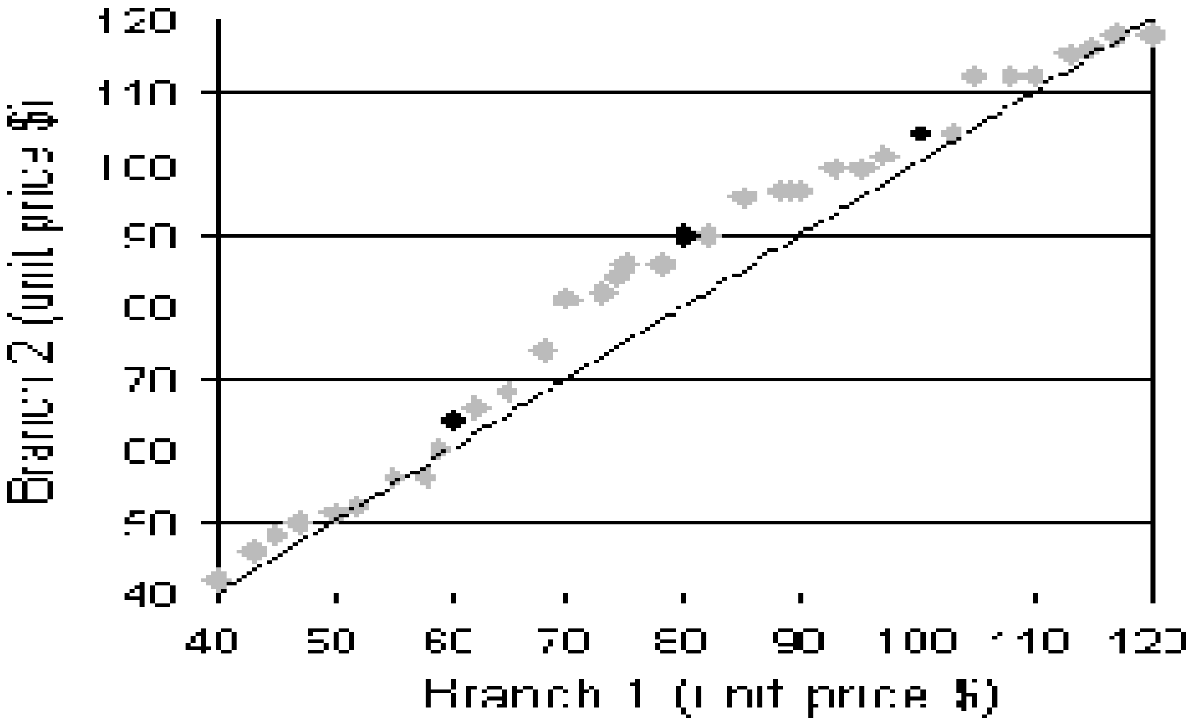
1. Quantile Plot

- Plots quantile information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value $x_{i.}$



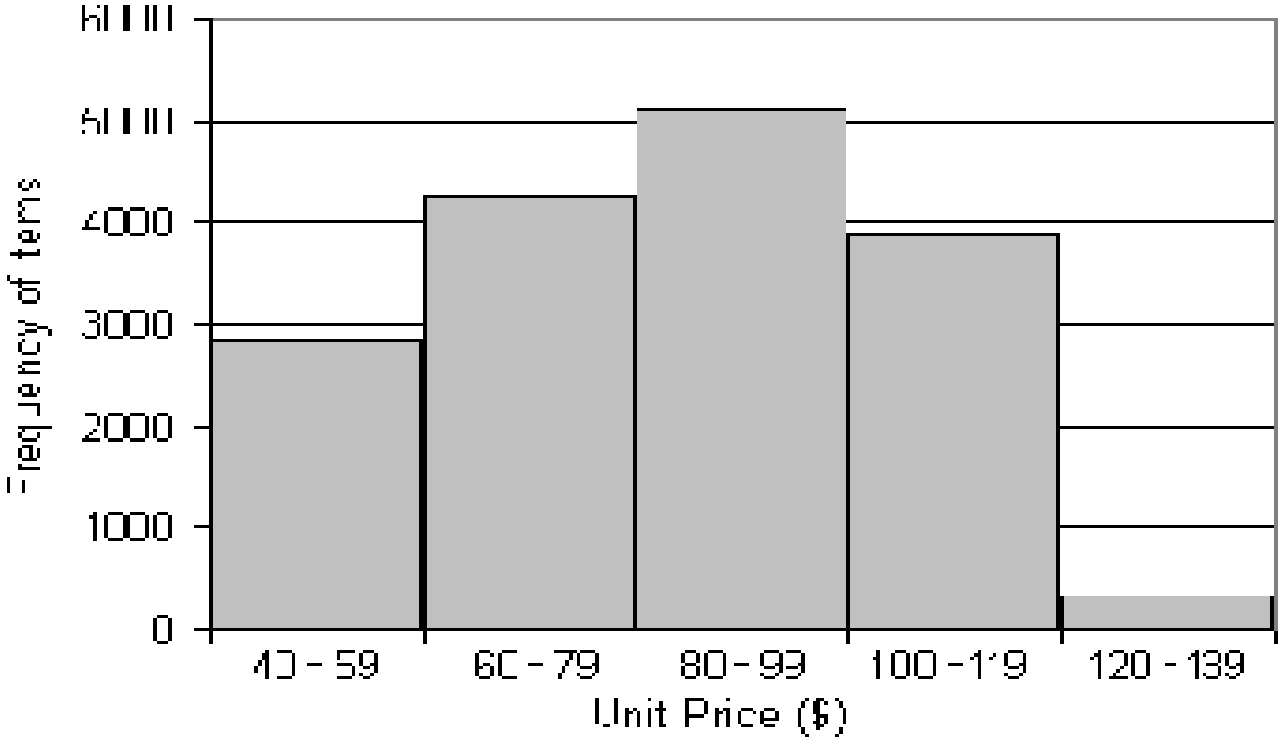
2. Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another.



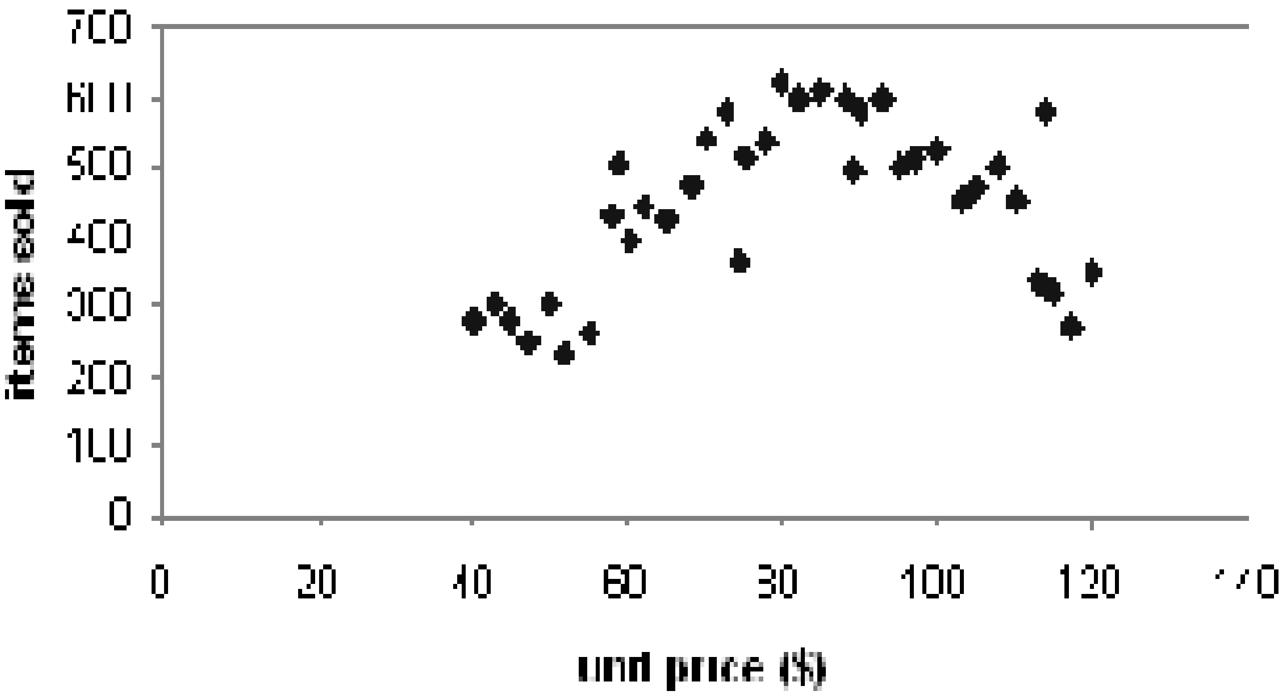
3. Histogram
(Frequency histograms)

- It Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data.

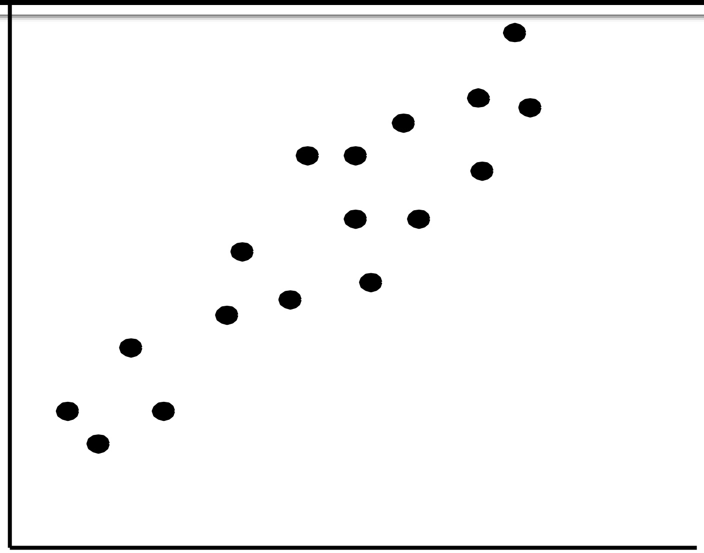


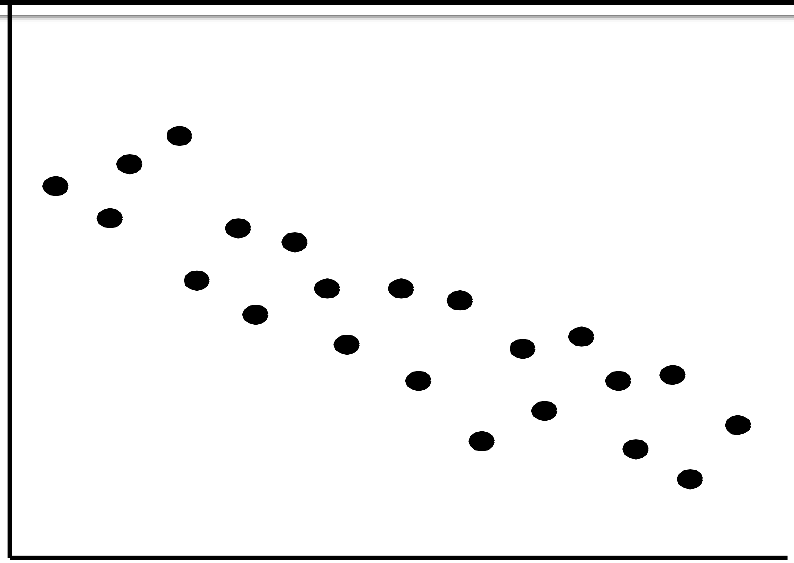
4. Scatter plot

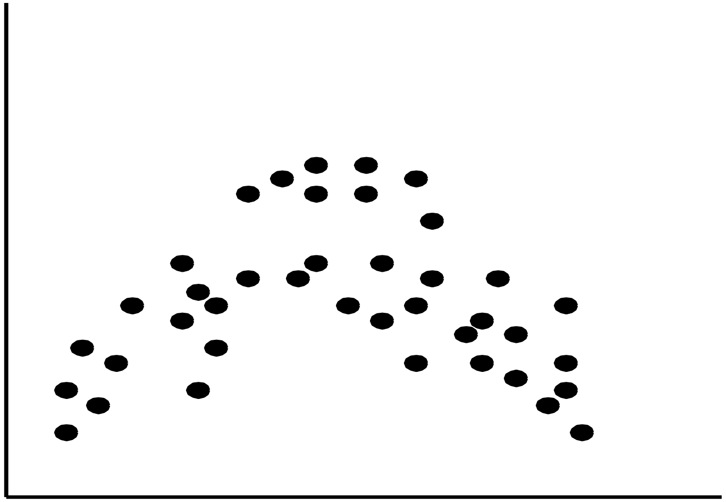
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data

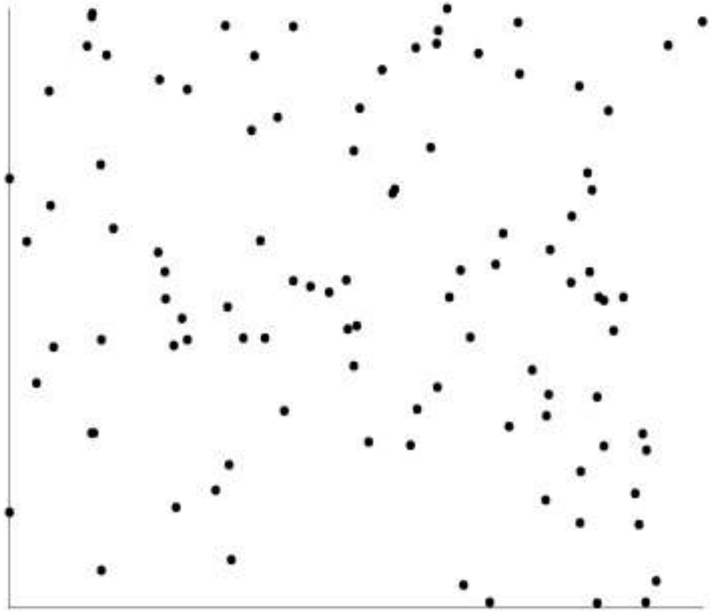


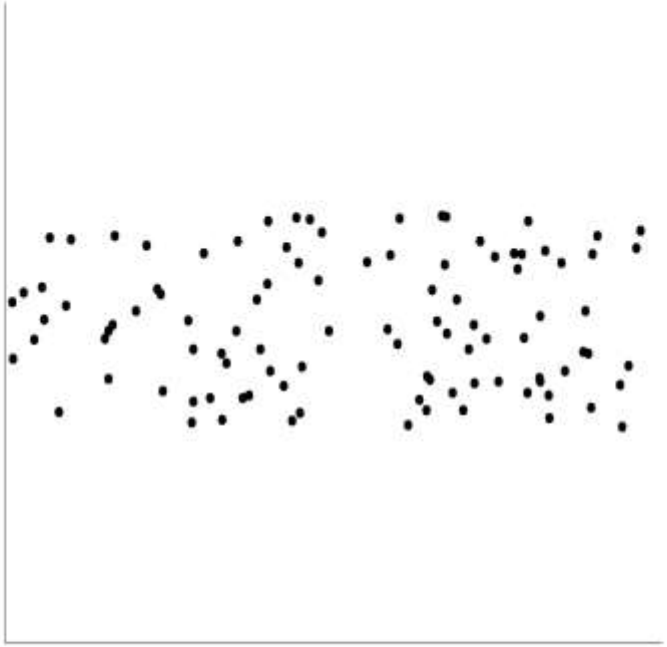


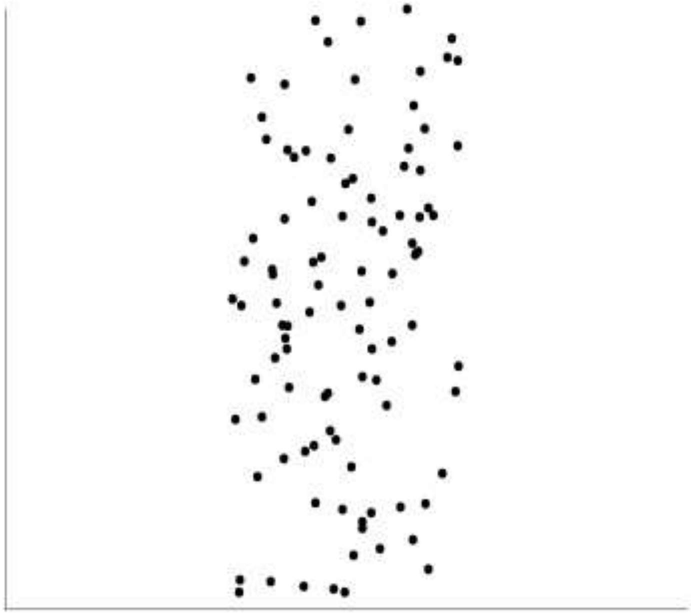


- The left half fragment is positively correlated
- The right half is negative correlated

Not Correlated Data







Data Cleaning as a Process

- **Discrepancy detection**
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections.
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- **Data migration and integration**
  - Data migration tools: allow transformations to be specified.
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface.
- **Integration of the two processes**
  - Iterative and interactive (e.g., Potter's Wheels).

Data Integration

Data Integration

Combines data from multiple sources into a coherent store.

- Entity identification problem:
  - **How to match data schema from different sources??**
  - Identify real world entities from multiple data sources, e.g., Customer_ID = Customer_Number
  - Possible reasons: different representations, different scales, e.g., metric vs. British units
  - Meta data can be used to avoid errors in schema integration.

Data Integration (cont'd)

- **Redundancy :** Redundant data occur often when One attribute can be derived from another attribute.

  - Redundancies can be detected by **Correlation Analysis.** It measures how strongly on attribute implies the other.

Correlation Analysis:

- **For Nominal Data: X^2 (chi-square) test**
- **For Numeric Data: Correlation analysis and covariance analysis.**

X² (chi-square) Correlation Test (For Nominal Data)

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The cells that contribute the most to the X² value are those whose actual count is very different from the expected count.

- The larger the X² value, the more likely the variables are related.

- Correlation does not imply causality as:
  - No of hospitals and No of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

Consider 2*2 contingency table data

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group.

Correlation Analysis (Numeric Data)

- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \overline{A})(b_i - \overline{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\overline{A}\overline{B}}{(n-1)\sigma_A \sigma_B}$$

Where, n is the number of tuples

$\overline{A}$ and $\overline{B}$ are the respective means of A and B,

σ_A and σ_B are the respective standard deviation of A and B, and

$\Sigma(a_i b_i)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{AB} < 0$: negatively correlated
- $r_{A,B} = 0$: independent;

Covariance Analysis (Numeric Data)

- Covariance is similar to correlation

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

Correlation coefficient: $$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

Where, n is the number of tuples,

$\overline{A}$ and $\overline{B}$ are the respective mean or **expected values** of A and B,

σ_A and σ_B are the respective standard deviation of A and B.

Covariance Analysis (cont'd)

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values.

- **Negative covariance**: If $Cov_{A,B} < 0$ then if one attribute is larger than its expected value, other is likely to be smaller than its expected value.

- **Independence**: $Cov_{A,B} = 0$ but the converse is not true:

Co-Variance: An Example

Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).

Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?

- $E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4$
- $E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6$
- $Cov(A,B) = (2\times5+3\times8+5\times10+4\times11+6\times14)/5 - 4 \times 9.6 =$ Thus, A and B rise together since Cov(A, B) > 0.

Data Reduction

Why data reduction?

- Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results.

Data Reduction Strategies

- Data cube aggregation

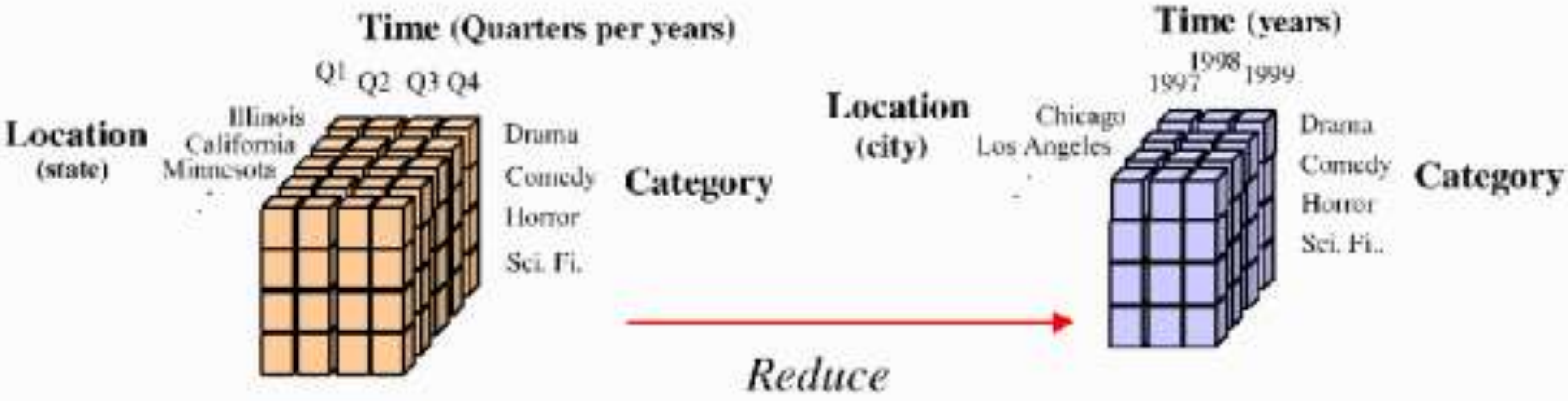
- Dimensionality reduction, e.g., remove unimportant attributes
  - Wavelet transforms
  - Principal Components Analysis (PCA)
  - Feature subset selection, feature creation

- Numerosity reduction
  - Regression, Histograms, clustering, sampling

- Decision Tree

1. Data Cube Aggregation

- **Reduce the data to the concept level needed in the analysis**
  - Use the smallest (most detailed) level necessary to solve the problem



- **Queries regarding aggregated information should be answered using data cube when possible**

Data Cube Aggregation

- **The aggregated data for an individual entity of interest**

  E.g., a customer in a phone calling data warehouse.

- **Multiple levels of aggregation in data cubes.**
  - Further reduce the size of data to deal with

- **Use the smallest representation which is enough to solve the task**

- **Queries regarding aggregated information should be answered using data cube, when possible**

2. Dimensionality Reduction

- Process of reducing number of attributes under consideration.

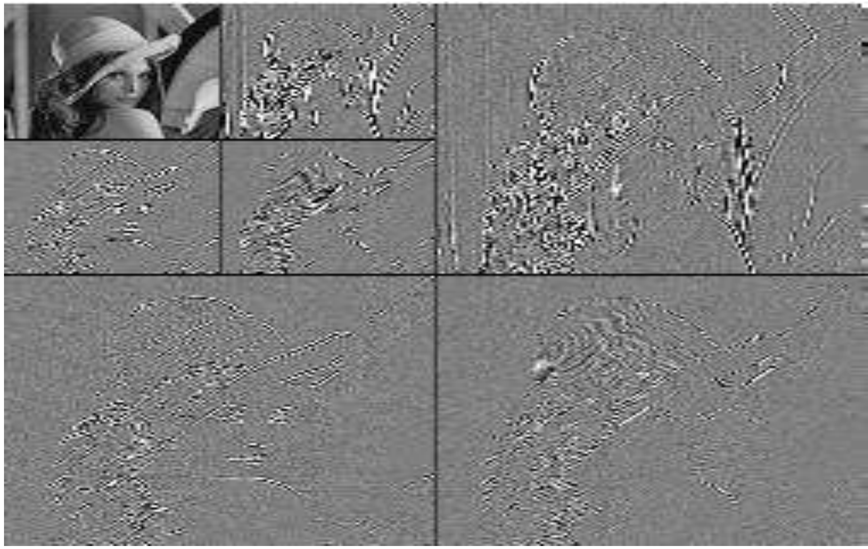
- **Includes:**
  - **Wavelet transform and PCA**, which transforms the original data onto a smaller space.
  - **Attribute subset selection**, which detect and remove irrelevant, weakly relevant or redundant attributes and dimensions.

Wavelet Transform

- Decomposes a signal into different frequency subbands
  - Applicable to n-dimensional signals also.

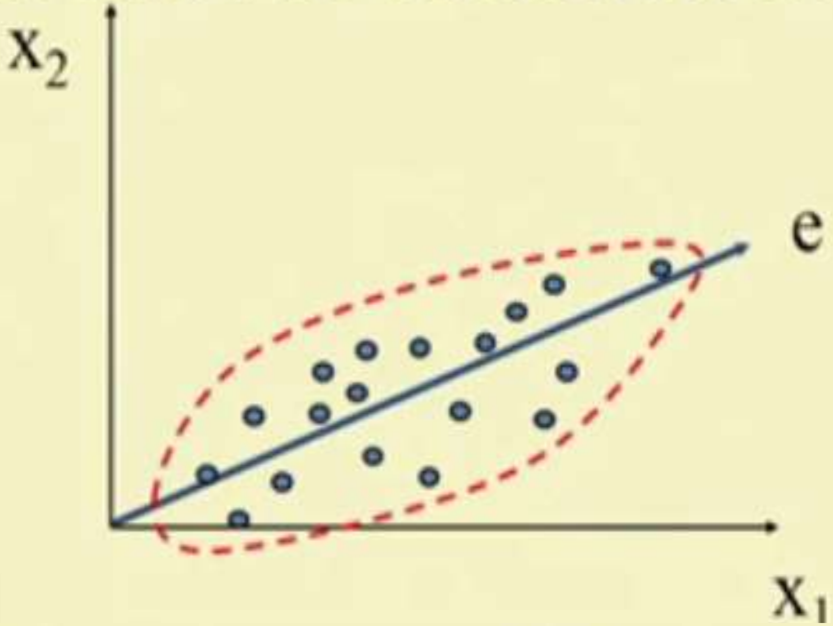
- Data are transformed to preserve relative distance between objects at different levels of resolution.

- Used for image compression.



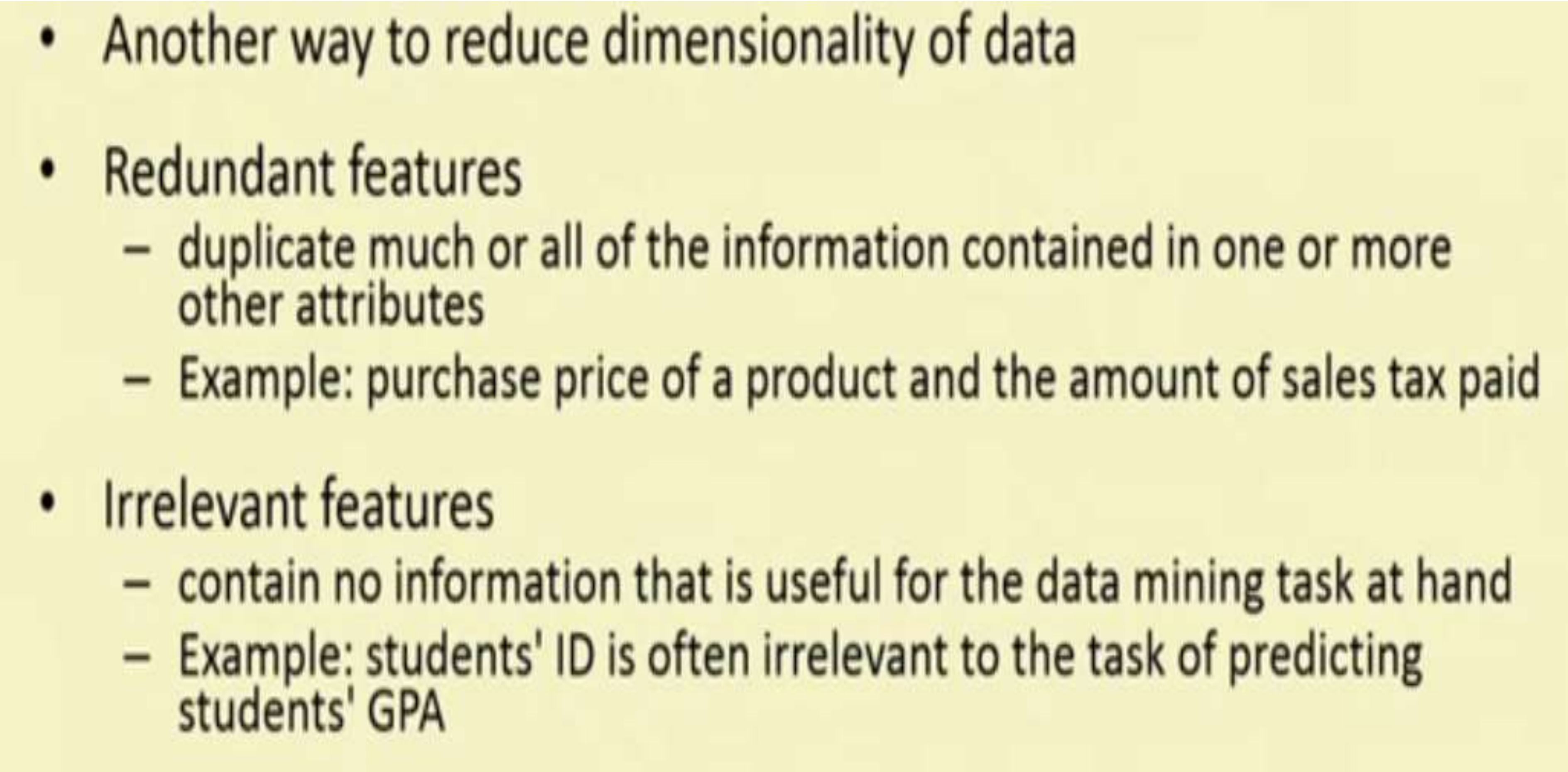
Principle Component Analysis

- Goal is to find a projection that captures the largest amount of variation in data



- Find the eigenvectors of the covariance matrix
- The eigenvectors define the new space

Attribute/ feature subset selection

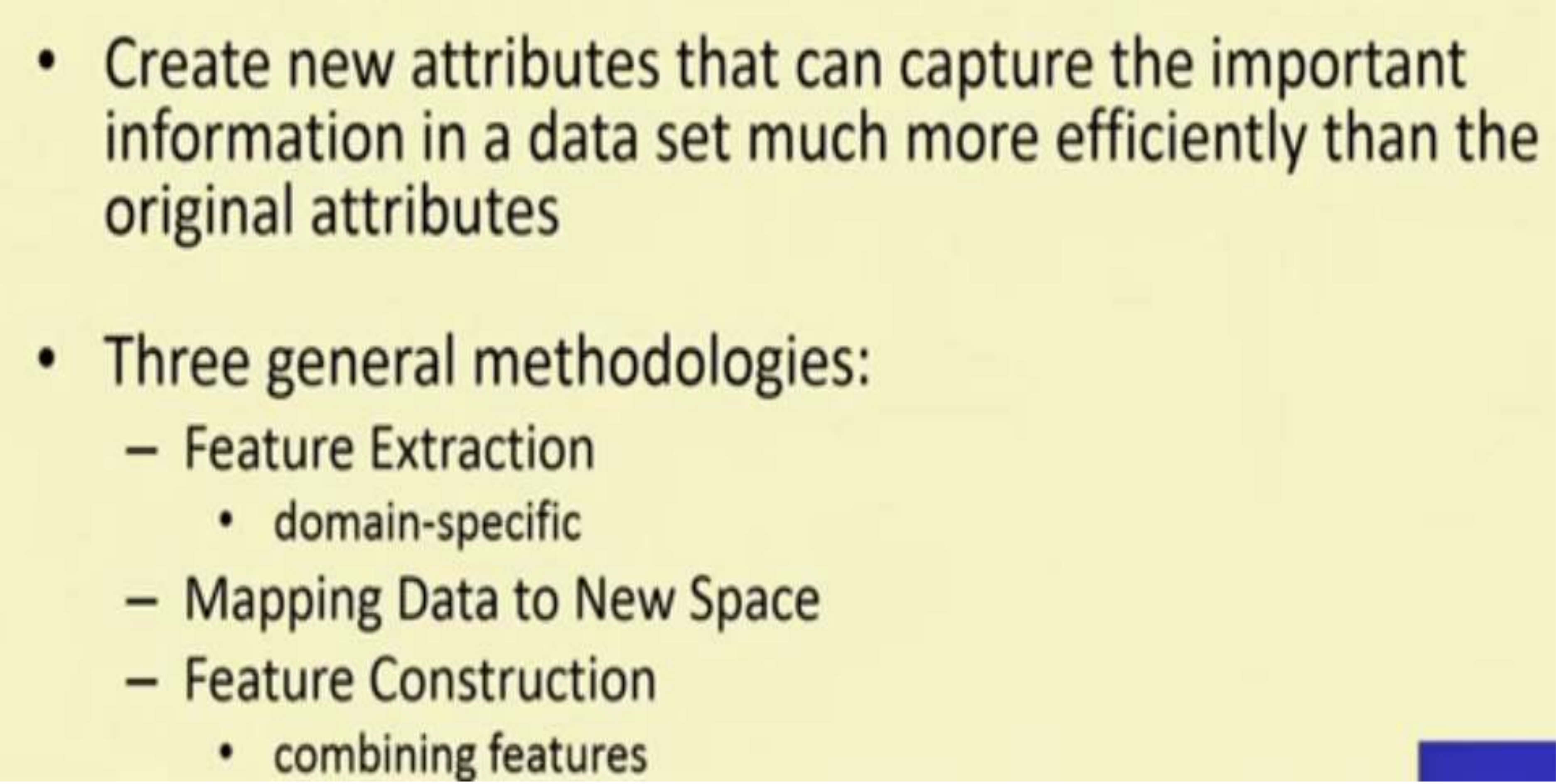


- Another way to reduce dimensionality of data
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

Attribute subset selection Techniques:

- Stepwise forward selection
- Stepwise backward elimination
- Combination of forward selection and backward elimination
- Decision tree induction.

Attribute/ Feature Creation



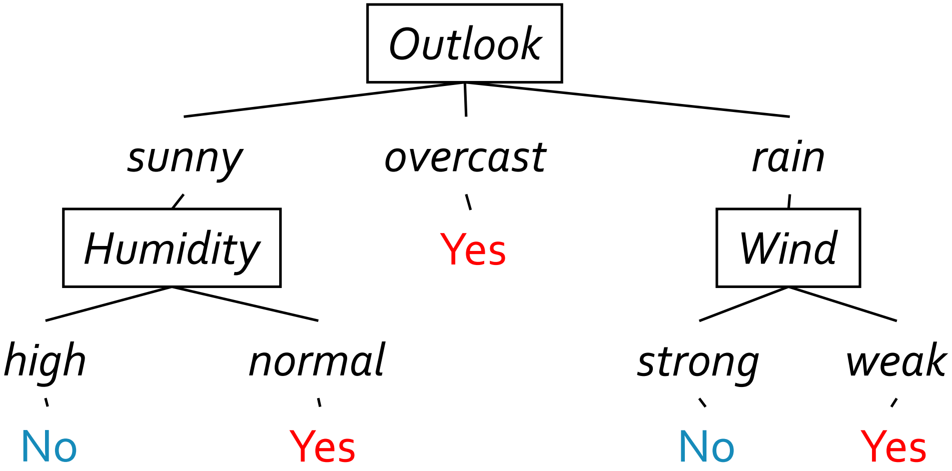
- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

- Three general methodologies:
  - Feature Extraction
    - domain-specific
  - Mapping Data to New Space
  - Feature Construction
    - combining features

Decision Tree

- Decision tree to represent learned target functions
  - Each internal node tests an attribute
  - Each branch corresponds to attribute value
  - Each leaf node assigns a classification

- Can be represented by logical formulas



Representation in decision trees

- Example of representing rule in DT's:
- *if* outlook = sunny AND humidity = normal
- OR
- *if* outlook = overcast
- OR
- *if* outlook = rain AND wind = weak
- *then* playtennis

3. Numerocity Reduction

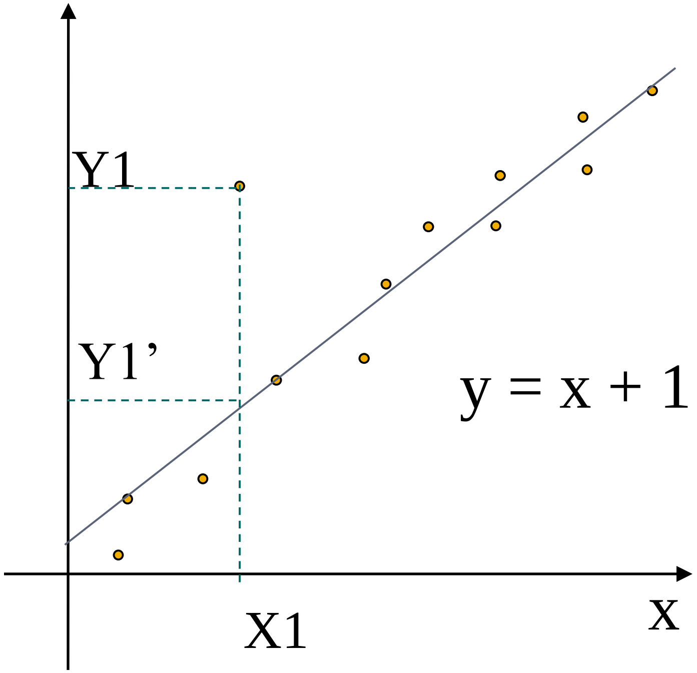
- Reduce data volume by choosing alternative, *smaller forms* of data representation
- **Parametric methods**
  - Assume the data fits some model, estimate model parameters, store only the parameters not data (except possible outliers)
  - Ex.: Log-linear models— estimate discrete *m*-D probability distribution.
- **Non-parametric methods**
  - Do not assume models, stores reduced representation of data
  - Includes histograms, clustering, sampling.

Regression Analysis

- Techniques for the modeling and analysis of numerical data consisting of values of a ***dependent variable*** (also called ***response variable (y)***) and of one or more *independent variables* (***explanatory variables*** or ***predictors (x)***).

- The parameters are estimated so as to give a "**best fit**" of the data

- Most commonly the best fit is evaluated by using the ***least squares method***.



- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Regress Analysis and Log-Linear Models

- Linear regression: $\boldsymbol{Y = w\,X + b}$
  - Two regression coefficients, *w* and *b*, specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of $Y_1, Y_2, \ldots, X_1, X_2, \ldots.$
- Multiple regression: $\boldsymbol{Y = b_0 + b_1 X_1 + b_2 X_2}$
  - Many nonlinear functions can be transformed into the above
- Log-linear models:
  - Approximate discrete multidimensional probability distributions
  - Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations
  - Useful for dimensionality reduction and data smoothing

Histogram

- A histogram for an attribute, A, partitions the data distribution of A into disjoint subsets (known as **buckets or bins**).

- Partitioning rules to determine buckets and attribute value partitioned:
  - **Equal-width histogram:** width of each bucket range is uniform.
  - **Equal-depth or equal-frequency histogram:** buckets are created so that frequency of each bucket is constant.

Clustering

Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only.

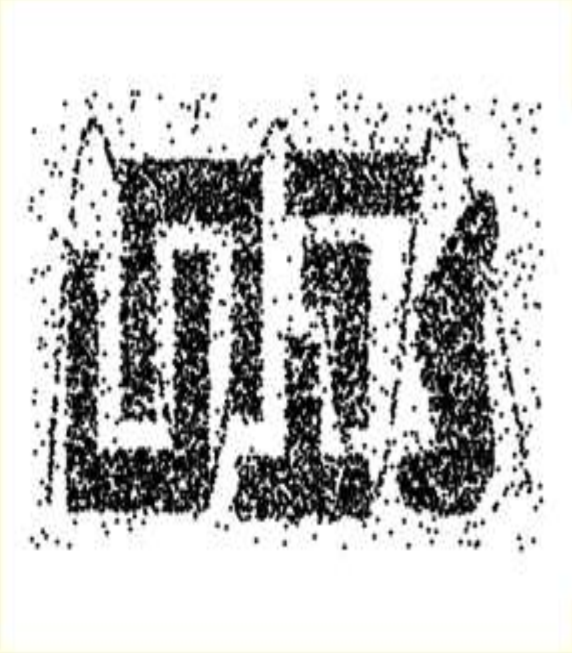
Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

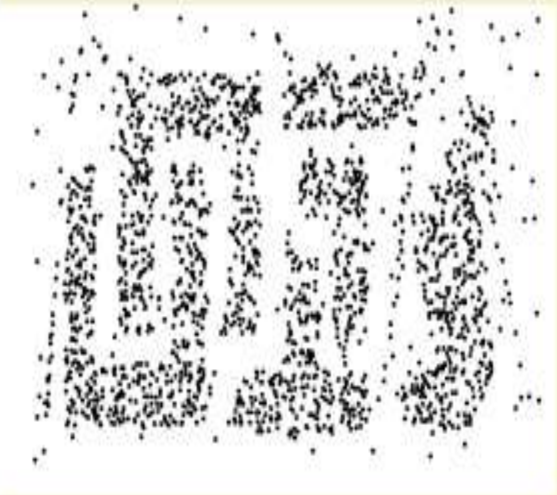
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

- The key principle for effective sampling :
  - using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

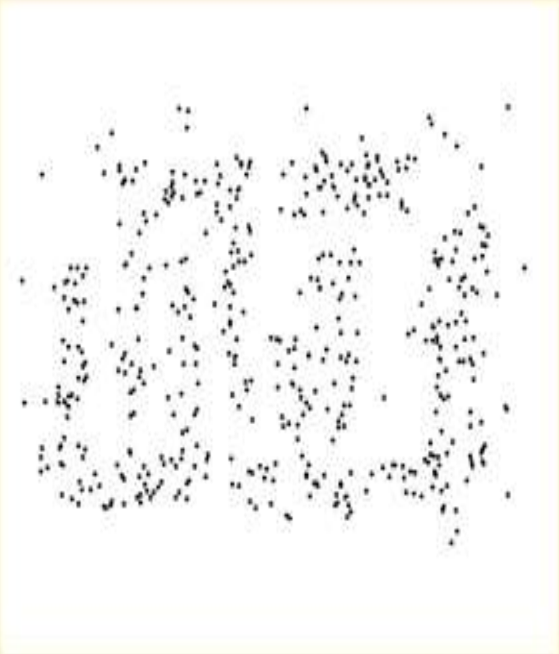
Sample Size



8000 points



2000 Points



500 Points

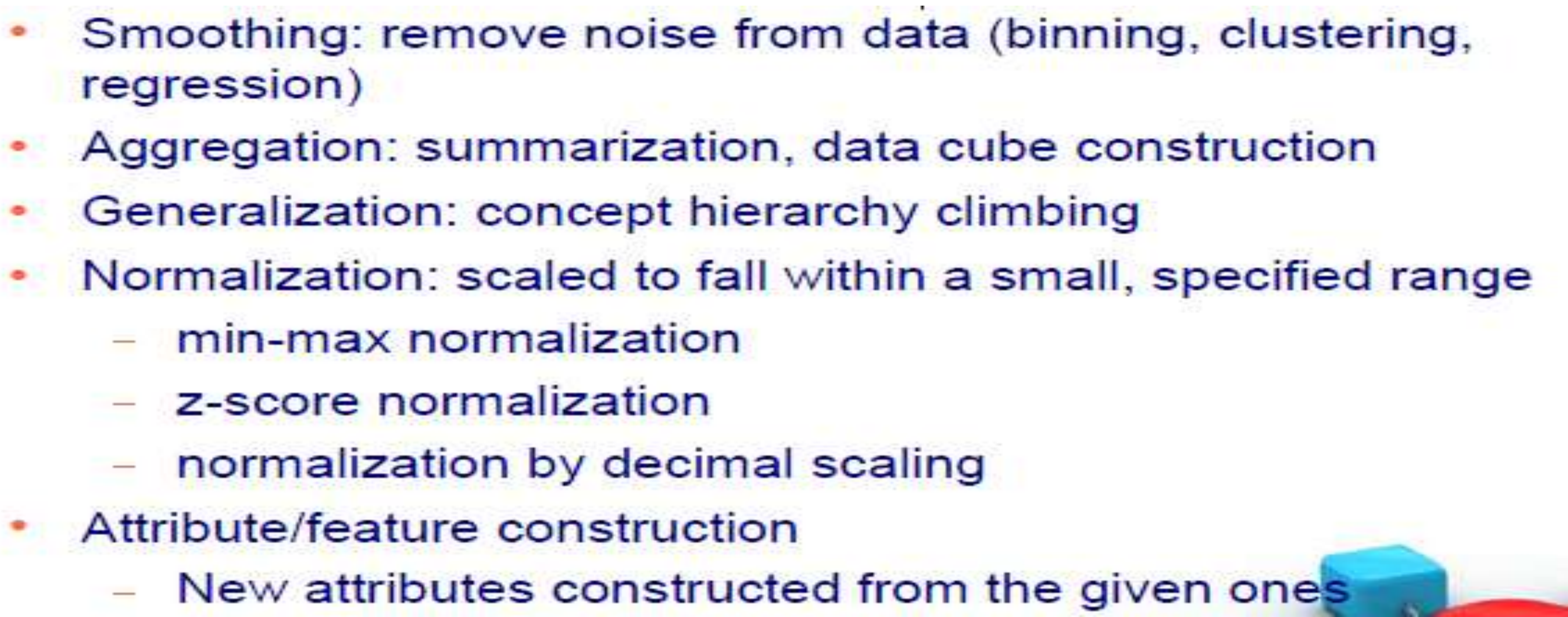
Types of Sampling

- Simple Random Sample without replacement (SRSWOR)
  - As each item is selected, it is removed from the population
- Simple Random Sample with replacement (SRSWR)
  - Objects are not removed from the population as they are selected for the sample. So, the same object can be picked up more than once
- Cluster Sample
  - Tuple in database are usually retrieved a page at a time, so that each page is considered as a cluster.
- Stratified sampling
  - Split the data into several partitions(strata); then draw random samples from each partition

Data Transformation

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values.
- **Methods:**



- Smoothing: remove noise from data (binning, clustering, regression)
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Ex. Let income range $12,000 to $98,000 normalized to [0.0, 1.0]. Then $73,000 is mapped to :

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max(|v'|) < 1

Data Discretization Methods

- Binning
  - Top-down split, unsupervised
- Histogram analysis
  - Top-down split, unsupervised
- Clustering analysis (unsupervised, top-down split or bottom-up merge)
- Decision-tree analysis (supervised, top-down split)
- Correlation (e.g., χ^2) analysis (unsupervised, bottom-up merge)

Binning Methods for Data Smoothing

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* **Partition into equal-frequency (equi-depth) bins:**
- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

* **Smoothing by bin means:**
- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

* **Smoothing by bin boundaries:**
- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

Concept Hierarchy Generation

- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity
- Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth*, *adult*, or *senior*)
- Concept hierarchy can be automatically formed for both numeric and nominal data. For numeric data, use discretization methods shown.

Concept Hierarchy Generation for Nominal Data

1. Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts.
   - *street* < *city* < *state* < *country*

2. Specification of a hierarchy for a set of values by explicit data grouping. E.g.{street, city} < state

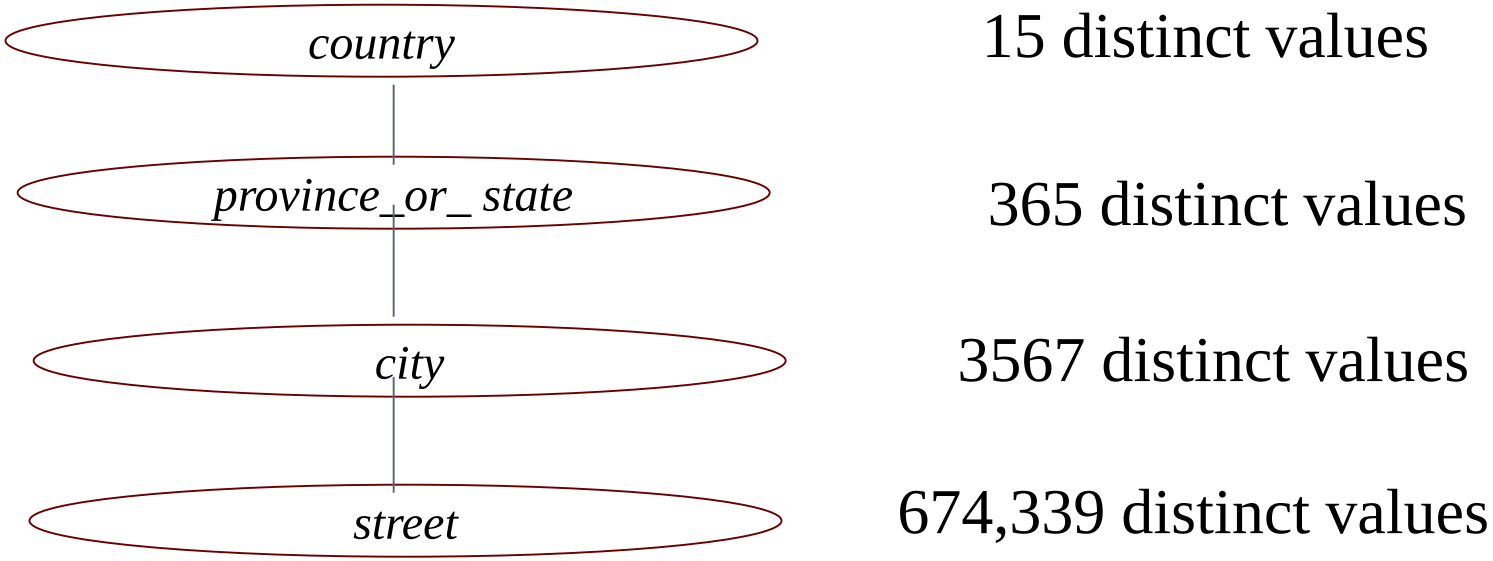
3. Specification of only a partial set of attributes.
   - E.g., only *street* < *city*, not others

4. Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values.
   - E.g., for a set of attributes: {*street, city, state, country*}

E.g. Automatic Concept Hierarchy Generation

Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set

- The attribute with the most distinct values is placed at the lowest level of the hierarchy
- Exceptions, e.g., weekday, month, quarter, year



country 15 distinct values

province_or_ state 365 distinct values

city 3567 distinct values

street 674,339 distinct values